# Quelques propositions pour la comparaison de partitions non strictes

Romain Quéré

UNIVERSITÉ DE LA ROCHELLE

*ÉCOLE DOCTORALE SCIENCES ET INGÉNIERIE POUR L'INFORMATION*

Laboratoire Mathématiques, Image et Applications

**THÈSE** présentée par :

**Romain QUÉRÉ**

**Le 6 Décembre 2012**

pour obtenir le grade de : **Docteur de l'université de La Rochelle**
Mention : **Automatique, Image et Signal**

# Quelques propositions pour la comparaison de partitions non strictes

---

**JURY :**

| | |
|---|---|
| **Gérard GOVAERT** | Professeur, Université de Technologie de Compiègne, Rapporteur |
| **Maria RIFQI** | Maître de Conférences HDR, Université Panthéon-Assas, Rapporteur |
| **Gilbert SAPORTA** | Professeur, Conservateur National des Arts et Métiers, Examinateur |
| **Michel BERTHIER** | Professeur, Université de La Rochelle, Examinateur |
| **Carl FRÉLICOT** | Professeur, Université de La Rochelle, Directeur de thèse |

# Résumé

Cette thèse est consacrée au problème de la comparaison de deux partitions non strictes (floues/probabilistes, possibilistes) d'un même ensemble d'individus en plusieurs clusters. Sa résolution repose sur la définition formelle de mesures de concordance reprenant les principes des mesures historiques développées pour la comparaison de partitions strictes et trouve son application dans des domaines variés tels que la biologie, le traitement d'images, la classification automatique.

Selon qu'elles s'attachent à observer les relations entre les individus décrites par chacune des partitions ou à quantifier les similitudes entre les clusters qui composent ces partitions, nous distinguons deux grandes familles de mesures pour lesquelles la notion même d'accord entre partitions diffère, et proposons d'en caractériser les représentants selon un même ensemble de propriétés formelles et informelles. De ce point de vue, les mesures sont aussi qualifiées selon la nature des partitions comparées. Une étude des multiples constructions sur lesquelles reposent les mesures de la littérature vient compléter notre taxonomie.

Nous proposons trois nouvelles mesures de comparaison non strictes tirant profit de l'état de l'art. La première est une extension d'une approche stricte tandis que les deux autres reposent sur des approches dite natives, l'une orientée individus, l'autre orientée clusters, spécifiquement conçues pour la comparaison de partitions non strictes. Nos propositions sont comparées à celles de la littérature selon un plan d'expérience choisi pour couvrir les divers aspects de la problématique. Les résultats présentés montrent l'intérêt des propositions pour le thème de recherche qu'est la comparaison de partitions. Enfin, nous ouvrons de nouvelles perpectives en proposant les prémisses d'un cadre qui unifie les principales mesures non strictes orientées individus.

**Mots-clés :** Comparaison de partitions, indice de Rand, indice de Jaccard, partition floue, partition possibiliste, cluster analysis, contingence-paires, matrice de contingence, matrice de coïncidence, norme triangulaire

# Abstract

This thesis is dedicated to the problem of comparing two soft (fuzzy/ probabilistic, possibilistic) partitions of a same set of individuals into several clusters. Its solution stands on the formal definition of concordance measures based on the principles of historical measures developped for comparing strict partitions and can be used in various fields such as biology, image processing and clustering.

Depending on whether they focus on the observation of the relations between the individuals described by each partition or on the quantization of the similarities between the clusters composing those partitions, we distinguish two main families for which the very notion of concordance between partitions differs, and we propose to characterize their representatives according to a same set of formal and informal properties. From that point of view, the measures are also qualified according to the nature of the compared partitions. A study of the multiple constructions on which the measures of the literature lie completes our taxonomy.

We propose three new soft comparison measures taking benefits of the state of art. The first one is an extension of a strict approach, while the two others lie on native approaches, one individual-wise oriented, the other cluster-wise, both specifically defined to compare soft partitions. Our propositions are compared to the existing measures of the literature according to a set of experimentations chosen to cover the various issues of the problem. The given results clearly show how relevant our measures are. Finally, we open new perspectives by proposing the premises of a new framework unifying most of the individual-wise oriented measures.

**Keywords :** Comparing partitions, Rand index, Jaccard index, fuzzy partition, possibilistic partition, cluster analysis, mismatch matrix, contingency matrix, coincidence matrix, triangular norm

# Remerciements

Je tiens à remercier l'ensemble du jury : Gérard Govaert et Maria Rifqi pour avoir accepté de rapporter cette thèse, ainsi que Gilbert Saporta pour avoir accepté d'examiner mes travaux.

Un grand merci à Carl Frélicot pour avoir dirigé cette thèse et s'y être tant impliqué, m'aiguillant et me conseillant bien plus que je n'aurais pu l'espérer durant ces trois années. Merci à lui encore pour ses qualités qui ont fait de la préparation de ce doctorat plus qu'une aventure scientifique, une aventure humaine.

Je remercie encore Michel Berthier pour avoir accepté d'examiner mon travail, mais aussi en sa qualité de directeur du laboratoire MIA et enfin pour les riches échanges partagés au Brussel's.

Aussi, je tiens à témoigner ma reconnaissance à Noël Fraisseix pour son aide précieuse et nos sympathiques discussions.

Enfin, merci à l'ensemble des membres du MIA pour leur accueil chaleureux. Je pense notamment à mes collègues doctorants passés et présents : Sloven, Agathe, José, Charles, Gina, ... ainsi qu'aux permanents qui, croisés au gré des couloirs, auront enjolivé mon quotidien : Christophe, Laurent, Renaud, Thierry, ... Une pensée aussi pour Sylviane, secrétaire du MIA, pour sa grande disponibilité et sa sympathie.

Plus personnellement, je souhaite remercier Caroline, pour sa patience à mon égard lors de la rédaction de ce mémoire, et tous les Tasdonnais de terre et d'adoption, sans lesquels cette tranche de vie n'aurait été aussi aussi riche en souvenirs. Remy, Mémère, Ado, Franaois, Pérou, Romain, Da et Damien, Max et Maxime, Joe et Pié : votre nom devrait désormais figurer quelque part à la Bibliothèque Nationale !

Bien sûr, je terminerai en remerciant ma famille et mes parents Christine et Michel, qui ont toujours su m'écouter, me soutenir et m'encourager.

# Sommaire

# Table des symboles

| Notation | Description | Page |
|---|---|---|
| $X$ | Ensemble de $n$ individus | 5 |
| $\mathbf{x}_k$ | $k$-ième individu de $X$ | 5 |
| $\{a, b\}$, $(a, b)$ | Paire, couple | 5 |
| $q$ | Nombre de paires dans $X$ | 18 |
| $\mathcal{C}(X)$ | Ensemble de tous les ensembles stricts définis sur $X$ | 32 |
| $\mathcal{F}(X)$ | Ensemble de tous les ensembles flous définis sur $X$ | 52 |
| $U$, $V$ | Matrice de $c$-partition, de $r$-partition de $X$ | 5 |
| $\mathbb{M}_{hcn}$ | Ensemble des $c$-partitions strictes | 7 |
| $\mathbb{M}_{fcn}$ | Ensemble des $c$-partitions floues | 7 |
| $\mathbb{M}_{pcn}$ | Ensemble des $c$-partitions possibilistes | 7 |
| $\mathbb{M}_{scn}$ | Ensemble des $c$-partitions non-strictes | 7 |
| $U^i$ | $i$-ème cluster de $U$, $i$-ème colonne de $U$ | 5 |
| $\mathbf{u}_k$ | Vecteur d'appartenance de $x_k$ aux classes de $U$, $k$-ème colonne de $U$ | 5 |
| $M(U, V)$ | Matrice de contingence-paires croisant $U$ et $V$, de termes $m_{\alpha\beta}$ | 20 |
| $m_{\alpha\beta}$ | Termes d'accord/désaccord négatifs/positifs, $\alpha, \beta \in \{0, 1\}$ | 20 |
| $N(U, V)$ | Matrice de contingence croisant $U$ et $V$, de terme général $n_{ij}$ | 28 |
| $n_{ij}$ | Cardinalité de l'ensemble $U^i \cap V^j$ | 28 |
| $\Psi_U$ | Matrice de coïncidence de $U$, de terme général $\psi_{U,kl}$ | 29 |
| $\psi_{U,kl}$ | Coïncidence des individus $\mathbf{x}_k$ et $\mathbf{x}_l$ selon la partition $U$ | 29 |
| $\Phi_U$ | Matrice de coïncidence normalisée de $U$, de terme général $\phi_{U,kl}$ | 85 |
| $I(U, V)$ | Mesure de concordance entre $U$, $V$, indice | 15 |
| $D(U, V)$ | Mesure de discordance entre $U$, $V$, distance | 15 |
| $RI(U, V)$ | Indice de Rand entre $U$, $V$ | 21 |
| $JI(U, V)$ | Indice de Jaccard entre $U$, $V$ | 21 |
| $\top$, $\perp$ | Norme, conorme triangulaire | 42 |
| $K_\top$ | Fonction normalisante | 42 |
| $\mathcal{I}_\top$ | Implication résiduelle | 42 |
| $\mathcal{S}$ | Mesure de sparsité | 42 |
| $\mathbb{1}_{mp}$ | Matrice $(m \times p)$ dont tous les éléments valent 1 | 123 |

# CHAPITRE 1 :

# Introduction générale

### Contexte et problématique

Les travaux que nous présentons concernent le problème de la caractérisation de l'accord que partagent deux points de vue, ou partitions, d'un même ensemble de données. Ce problème naturel se rencontre dans de nombreux domaines (biologie [Chavent et al., 2001], psychologie et sciences sociales [Arabie et Boorman, 1973], gestion et management [Charon et al., 2007], recherche de consensus en classification automatique [Krieger et Green, 1999], évaluation de méthodes de segmentation ou d'espaces colorimétriques en traitement d'images [Unnikrishnan et al., 2007], etc) et revêt plusieurs aspects. Lorsque l'on cherche à extraire de l'information au sein d'un ensemble de données, une méthodologie récurrente consiste à regrouper les éléments de l'ensemble selon des caractéristiques communes. Dans le cas où ces données sont étiquetées selon une vérité terrain dont on dispose, la recherche de la méthode d'évaluation ou celle des caractéristiques qui ont mené à un tel étiquetage repose généralement sur une approche exploratoire consistant à produire plusieurs partitions dont la concordance avec l'étiquetage à disposition est évaluée. Dans d'autres cas, aucune information n'est disponible a priori sur les données et on cherche alors à mesurer l'accord entre les partitionnements réalisés par plusieurs évaluateurs (humains, algorithmes) afin d'avérer des structures de groupe communes et les caractériser. Enfin, on peut a contrario chercher à minimiser la concordance entre deux partitions afin que celles-ci se complètent de sorte de fournir une description la plus complète possible des données qu'elles caractérisent. La résolution de ces problèmes repose sur la définition formelle de mesures de comparaison de deux partitions, et un effort théorique et pratique important a été porté en ce sens depuis la deuxième moitié du siècle dernier.

Le cadre mathématique utilisé pour la modélisation des partitions a été élargi afin d'autoriser les éléments à appartenir aux groupes de manière partielle (selon des

principes d'imprécision et/ou d'incertitude), et a mené à l'apparition de partitions dites non déterministes. Ces partitions offrent une plus grande souplesse pour l'analyse des données qu'elles caractérisent, et c'est la raison pour laquelle nombre de méthodes modernes de fouille de données s'attachent aujourd'hui à produire des partitions de ce type. La définition d'une nouvelle génération de mesures de comparaison adaptées à ce cadre élargi est donc nécessaire, et cette nécessité constitue le fondement de cette thèse.

**Plan du mémoire**

Ce mémoire est organisé en cinq chapitres dont cette introduction générale en est le premier.

Le second chapitre (page 5) est consacré à la définition des enjeux et concepts clés de la problématique de la comparaison de partitions, et se termine par une revue des mesures de comparaison de partitions déterministes, aujourd'hui considérées comme historiques car les premiers travaux remontent aux années soixante-dix et parce que ces propositions posent les bases de nombre de mesures de comparaison non déterministes. Cette revue est menée selon une approche taxonomique identifiant les caractéristiques propres à chacune de mesures présentées. Ces mesures sont systématiquement illustrées par les mêmes exemples afin de mettre en évidence les concepts qu'elles induisent et les mécanismes de calculs sur lesquels elles reposent.

Le troisième chapitre (page 41) dresse un état de l'art le plus complet possible de la comparaison de partitions non déterministes. Nous y reprenons la même taxonomie que celle adoptée au deuxième chapitre et nous explicitons certains liens existant entre les propositions présentées. Encore une fois, des exemples sont systématiquement donnés pour chacunes d'entre-elles afin d'en illustrer les propriétés. Ce chapitre se conclut par deux discussions. La première concerne les complexités en temps et en espace des mesures revues, tandis que la seconde concerne leur capacité à comparer des partitions de nature différentes.

Dans le quatrième chapitre (page 79), nous décrivons nos contributions au domaine de la comparaison de partitions non déterministes. Nous y présentons trois nouveaux indices fondés chacun sur une construction différente ([Quéré et al., 2010; Quéré et Frélicot, 2011a, 2012]). Une même série d'expérimentations a été menée de manière à évaluer le comportement de chacun d'entre eux et les comparer à des mesures revues au chapitre précédent. Nous proposons aussi à la fin de ce chapitre un formalisme préparant la définition d'un cadre unifiant une partie des mesures de

comparaison non déterministes, à la lumière duquel sont étudiées ces dernières.

Enfin, le cinquième chapitre (page 131) est une conclusion générale où sont tout d'abord résumés nos travaux sous l'angle des apports scientifiques qu'ils constituent. Quelques perspectives concernant de futurs travaux y sont ensuite abordées.

# CHAPITRE 2 :

## Définitions, historique et enjeux de la comparaison de partitions

**Résumé :** *Ce chapitre préliminaire a pour objectif de dresser le cadre général dans lequel s'inscrit la problématique de la comparaison de partitions. Après en avoir introduit les notions fondamentales, il se conclut par une revue des mesures strictes historiques, menée selon une approche taxonomique.*

### 2.1 Qu'est-ce qu'une partition ?

Une partition d'un ensemble de données en est une description structurelle particulière, quel que soit leur type (données homogènes ou non, quantitatives, qualitatives, etc). Elle décrit une structure de groupes réunissant des éléments qui partagent les mêmes caractéristiques. Ces groupes sont appelés *classes* ou *clusters* ci-après selon qu'ils sont le résultat d'un étiquetage par un expert ou par un algorithme de *partitionnement* ou *clustering*[1]. Leur union doit couvrir l'ensemble complet des données, et s'ils sont disjoints, la partition est dite *stricte*[2].

**Définition 1.** *Une* partition stricte $P_U = \{P_{U^1}, P_{U^2}, ..., P_{U^c}\}$ *d'un ensemble* $X$ *est un ensemble de parties non-vides et disjointes de* $X$*, couvrant* $X$ *:*

- $P_{U^i} \neq \emptyset \ \forall i \in \{1, 2, ..., c\}$,
- $P_{U^i} \cap P_{U^j} = \emptyset \ \forall i, j \in \{1, 2, ..., c\}$,
- $\bigcup_{i=1}^{c} P_{U^i} = X$.

1. On prendra pour convention de réserver l'usage du terme de *classe* pour évoquer spécifiquement les sous-ensembles d'une partition experte, préférant ainsi utiliser celui de *cluster* lorsque la distinction n'est pas de mise.

2. Les mathématiciens parlent de *partition* là où notre communauté adjoint ce qualificatif pour distinguer ce type de partition d'autres types décrits plus loin ; pour les partitions non strictes, ils utilisent le terme *recouvrement*.

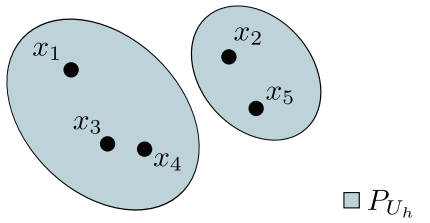



FIGURE 2.1 – Partition stricte $P_{U_h}$ de données $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ en deux clusters

On dira que $P_U$ est une $c$-partition de $X$ si on souhaite préciser le nombre de groupes qu'elle décrit. Si l'ensemble de données est composé de $n$ éléments (ou *individus*), c'est-à-dire que $X = \{\mathbf{x}_1,\ \mathbf{x}_2,\ ...,\ \mathbf{x}_n\}$, alors, à toute $c$-partition $P_U$ de $X$, on peut faire correspondre une *matrice de partition* $U$ de taille (c × $n$)

$$U = \begin{array}{c} \begin{array}{cccccc} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k & \cdots & \mathbf{u}_n \end{array} \\ \left( \begin{array}{cccccc} u_{11} & u_{12} & \cdots & u_{1k} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2k} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{i1} & \cdots & \cdots & u_{ik} & \cdots & u_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{c1} & u_{c2} & \cdots & u_{ck} & \cdots & u_{cn} \end{array} \right) \end{array} \begin{array}{c} \\ U^1 \\ U^2 \\ \vdots \\ U^i \\ \vdots \\ U^c \end{array}$$

dont le terme général $u_{ik}$ représente le degré selon lequel le $k$-ième individu ($k = 1, n$) de $X$ appartient au $i$-ième cluster de $P_U$ ($i = 1, c$), binaire pour une partition stricte, et où chacune des colonnes $\mathbf{u}_k$, appelées *vecteurs d'appartenance*, regroupe les degrés d'appartenance de chaque individu à chacun des clusters de $P_U$. Par simplicité et confort de notation, on se permettra par la suite d'assimiler abusivement une partition $P_U$ à sa matrice de partition $U$, de sorte que la partition sera définie à une permutation des lignes de sa matrice de partition près. *In extenso*, on se réfèrera aussi aux lignes $U^i$ de la matrice $U$ pour évoquer les clusters de cette même partition.

**Exemple 1.** *Partition stricte $P_{U_h}$ et sa matrice de partition $U_h$ selon le partitionnement de l'ensemble d'individus $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ en deux clusters tel qu'illustré par la Figure 2.1.*

$$P_{U_h} = \{\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4\}, \{\mathbf{x}_2, \mathbf{x}_5\}\}$$

$$U_h = \begin{array}{c} \begin{array}{ccccc} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_4 & \mathbf{u}_5 \end{array} \\ \left( \begin{array}{ccccc} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right) \end{array} \begin{array}{c} \\ U_h^1 \\ U_h^2 \end{array}$$

Il existe d'autres types de partitions que les partitions strictes. Selon que les clusters qui la composent sont ou non mutuellement exclusifs et selon que les individus qu'elle caractérise peuvent ou non n'appartenir que partiellement à chaque cluster, une partition peut être modélisée par différentes théories mathématiques desquelles découlent différents espaces de partitions. On recense ainsi en plus de celui des partitions strictes les deux principaux espaces que sont ceux des partitions *floues* (ou *probabilistes*) et *possibilistes*, définis dans Bezdek [1981], respectivement par :

- $\mathbb{M}_{pcn} = \{U \in \mathbb{R}^{c\,n} : u_{ik} \in [0,1]\}$,
- $\mathbb{M}_{fcn} = \left\{U \in \mathbb{M}_{pcn} : \sum_{i=1}^{c} u_{ik} = 1\right\}$.

Dans un souci d'homogénéité, on redéfinit de la même façon l'espace des partitions strictes :

- $\mathbb{M}_{hcn} = \{U \in \mathbb{M}_{fcn} : u_{ik} \in \{0,1\}\}$.

Ces ensembles sont imbriqués : $\mathbb{M}_{hcn} \subset \mathbb{M}_{fcn} \subset \mathbb{M}_{pcn}$. Par commodité, on retiendra aussi l'ensemble des partitions non-strictes (ou *soft partitions*) tel que défini dans [Anderson et al., 2010] :

- $\mathbb{M}_{scn} = \mathbb{M}_{pcn} \ominus \mathbb{M}_{hcn}$, où $\ominus$ symbolise la soustraction ensembliste,

auquel nous ajoutons les restrictions de partitions possibilistes suivants :

- $\mathbb{M}^{>}_{pcn} = \left\{U \in \mathbb{M}_{pcn} : \exists\, k \in \{1,\dots,n\}, \sum_{i=1}^{c} u_{ik} > 1\right\}$,
- $\mathbb{M}^{\leq}_{pcn} = \mathbb{M}_{pcn} \ominus \mathbb{M}^{>}_{pcn}$

à partir desquels on peut définir les restrictions de partitions non strictes $\mathbb{M}^{>}_{scn}$ et $\mathbb{M}^{\leq}_{scn}$ correspondantes. Des partitions appartenant à certains de ces ensembles sont données à l'Exemple 2.

**Exemple 2.**

$$
\begin{array}{rc}
 & \underline{\textit{partitions :}} \\
U_h = & \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \\
U_f = & \begin{pmatrix} 0.9 & 0.2 & 0.7 & 0.6 & 0.3 \\ 0.1 & 0.8 & 0.3 & 0.4 & 0.7 \end{pmatrix} \\
U_{p^{>}} = & \begin{pmatrix} 0.9 & 0.2 & 0.7 & 0.6 & 0.3 \\ 0.3 & 0.8 & 0.3 & 0.4 & 0.7 \end{pmatrix} \\
U_{p^{\leq}} = & \begin{pmatrix} 0.5 & 0.2 & 0.7 & 0.6 & 0.3 \\ 0.3 & 0.8 & 0.3 & 0.4 & 0.7 \end{pmatrix}
\end{array}
$$

| $\mathbb{M}_{h25}$ | $\mathbb{M}_{f25}$ | $\mathbb{M}_{p25}$ | $\mathbb{M}_{s25}$ | $\mathbb{M}^{>}_{p25}$ | $\mathbb{M}^{\leq}_{s25}$ |
|---|---|---|---|---|---|
| • | • | • | | | |
| | • | • | • | | • |
| | | • | • | • | |
| | | • | • | | • |

Étant donnée une partition non stricte, on peut facilement lui associer une partition stricte.

**Définition 2.** *La défuzzification par maximum d'appartenance est l'opération :* $\mathbb{M}_{scn} \longrightarrow \mathbb{M}_{hcn}, U \mapsto U'$ *telle que,* $\forall k \in \{1, 2, ..., n\}$

$$u'_{ik} = \begin{cases} 1 & si\ i = argmax_{j=1,c} u_{jk} \\ 0 & sinon \end{cases} . \tag{2.1}$$

*On dira que* $U'$ *est la partiton* $U$ *max-défuzzifiée.*

**Définition 3.** *Deux partitions sont dites max-compatibles si leur défuzzification par maximum d'appartenance sont égales.*

Par définition, toute partition de $\mathbb{M}_{scn}$ est bien évidemment max-compatible avec sa partition max-défuzzifiée.

**Exemple 3.** *Les partitions non strictes de l'Exemple 2 sont toutes max-compatibles.*

Chacun des trois ensembles $\mathbb{M}_{hcn}$, $\mathbb{M}_{fcn}$ et $\mathbb{M}_{pcn}$ porte une sémantique propre qui doit être précisée. Ainsi les partitions strictes, reposant sur la traditionnelle théorie des ensembles, sont des partitions déterministes pour lesquelles l'appartenance d'un individu à un cluster donné est totale et exclusive ; les clusters sont donc disjoints deux à deux et un individu ne peut appartenir qu'à un et un seul cluster, si bien que leur interprétation reste triviale. En autorisant une appartenance partielle des individus aux clusters, les partitions de $\mathbb{M}_{fcn}$ permettent quant à elles de modéliser :

- soit l'incertitude de l'appartenance dans le cas des partitions probabilistes, pour lesquelles les degrés d'appartenance doivent être interprétés comme des probabilités d'appartenance *a posteriori* des individus aux clusters,
- soit l'imprécision sur l'appartenance, voire selon certains auteurs la vraisemblance de cette même appartenance [Gath et Geva, 1989] dans le cas des partitions floues.

Toutefois, la contrainte de normalisation imposée aux vecteurs d'appartenance amène à ne devoir considérer chaque degré que relativement aux autres pour un vecteur d'appartenance donné, si bien que les partitions floues ne permettent pas de mettre aisément en évidence les individus atypiques au sein de la structure décrite. C'est sur la base de cette critique que repose la construction des partitions possibilistes [Krishnapuram et Keller, 1993], qui propose de considérer individuellement, et donc de manière absolue, l'appartenance des individus à chacun des clusters, modélisant ainsi le degré de typicalité des individus à ces derniers, soit à la fois l'incertitude et l'imprécision. Il est donc important de garder en mémoire que, bien qu'il existe une certaine

classe de partitions possibilistes satisfaisant les contraintes imposées aux partitions floues puisque l'ensemble de ces dernières est imbriqué dans celui des partitions possibilistes, les éléments la constituant ne doivent en aucun cas être interprétés sous le prisme de la sémantique propre aux partitions floues.

**Définition 4.** *[Zadeh, 1965] Un ensemble flou $U^i$ est inclus dans un autre ensemble flou $V^j$ si :*

$$u_{ik} \leq v_{jk},\ \forall k \in \{1, 2, ..., n\}. \tag{2.2}$$

*On note une telle relation $U^i \subseteq V^j$.*

Rien n'empêche d'étendre cette définition aux lignes de n'importe quelles matrices de partition de $\mathbb{M}_{pcn}$, considérées comme des ensembles stricts, flous ou possibilistes, de sorte que $U^i \subseteq V^j$ symbolise l'inclusion de clusters.

**Définition 5.** *Une c-partition $U$ est un raffinement d'une r-partition $V$ si et seulement si :*

- $c \geq r$,
- $\forall i \in \{1, 2, ..., c\}, \exists j$ *tel que* $U^i \subseteq V^j$.

*$U$ est alors dite plus fine que $V$ qui elle-même est dite plus grossière que $U$, et on note cette relation $U \subseteq V$.*
*Si $c > r$, alors le raffinement est dit* strict, *et on notera $U \subset V$.*

Autrement dit, une partition est plus fine qu'une autre si ses clusters sont inclus dans ceux de l'autre. Dans le cas particulier de partitions de même dimension, c'est-à-dire ayant le même nombre de clusters ($c = r$), il est évident qu'on peut construire autant de partitions $V$ plus fines qu'une partition $U$ de $\mathbb{M}_{pcn}$ que l'on veut. Il suffit de prendre $V = \alpha U$, avec $\alpha \in ]0, 1]$. Des exemples de raffinements stricts sont donnés à l'Exemple 4.

**Exemple 4.**

- $U_h = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ *est plus fine que* $V_h = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
- $U_f = \begin{pmatrix} 0.9 & 0.2 & 0 & 0 \\ 0.1 & 0.7 & 0 & 0.5 \\ 0 & 0.1 & 0.8 & 0 \\ 0 & 0 & 0.2 & 0.5 \end{pmatrix}$ *est plus fine que* $V_f = \begin{pmatrix} 0.9 & 0.2 & 0 & 0 \\ 0.1 & 0.7 & 0.8 & 0.5 \\ 0 & 0.1 & 0.2 & 0.5 \end{pmatrix}$

– $U_p = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0 & 0.6 & 0.8 & 0.4 \\ 0.4 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.3 \end{pmatrix}$ *est plus fine que* $U_f$ *et* $V_f$.

Terminons cette Section sur quelques propriétés particulières propres aux partitions strictes de $\mathbb{M}_{hcn}$. Contrairement aux partitions provenant des autres espaces, l'ensemble des partitions de $\mathbb{M}_{hcn}$ d'un ensemble $X$ d'invididus est fini. Cette particularité lui confère la propriété de définir un treillis[3], selon la relation d'ordre partiel donnée par le raffinement strict des partitions qui le constituent. Les bornes supérieure et inférieure du treillis sont respectivement les deux partitions particulières que sont la 1-partition, ou partition singletons, $\mathbb{1}_{1n}$ = (1 1 ... 1), regroupant tous les individus au sein d'un seul et unique cluster, et la partition identité $I_n$ constituée uniquement de clusters singletons. Notons toutefois que ces deux cas limite de partitions ne traduisent bien évidemment aucune structure de groupes au sein de l'ensemble $X$ qu'ils caractérisent. Le diagramme de Hasse présenté à la Figure 2.2 illustre le treillis défini sur l'ensemble des partitions de l'ensemble des quatre premiers entiers $\{1,2,3,4\}$. La relation d'ordre se lit de bas en haut, *e.g.* l'arc entre la partition singletons $\mathbf{1}_4 = \{\{1\},\{2\},\{3\},\{4\}\}$ et la 3-partition $\{\{1\},\{2,3\},\{4\}\}$ indique que la première est un raffinement strict de la seconde.

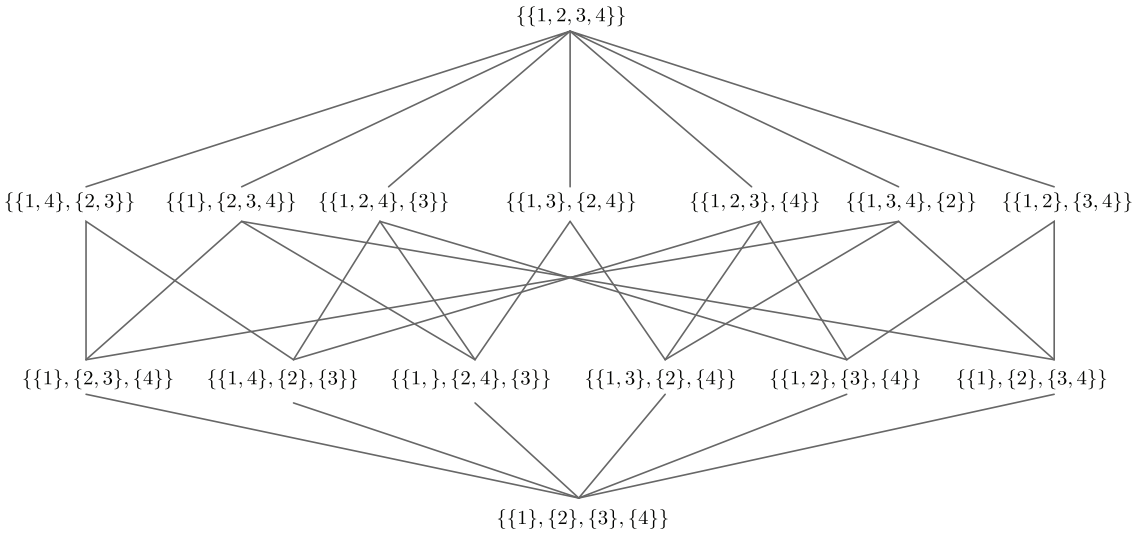



FIGURE 2.2 – Diagramme de Hasse du treillis des partitions strictes de l'ensemble $\{1,2,3,4\}$, ordonnées selon leur raffinement.

3. Un treillis est un ensemble E muni d'une relation d'ordre $\subseteq$ vérifiant :
$\forall\, a, b \in E$, $\exists$ une borne supérieure et une borne inférieure à l'ensemble {a,b}.

## 2.2 De l'origine des partitions

Une partition peut-être produite de plusieurs façon. Elle peut être issue d'un étiquetage manuel des données par un sujet humain, expert ou non du domaine concerné, et décrit alors :

- soit une absolue vérité-terrain – on parle alors de *partition-expert*,
- soit un partitionnement subjectif, que l'on cherchera généralement alors à confronter à d'autres partitions du même type.

Dans ces deux cas, une partition prend la plupart du temps ses valeurs dans $\mathbb{M}_{hcn}$. Une partition peut aussi être le résultat d'une méthode ou d'un algorithme de classification automatique non-hiérarichique ou *clustering*, dont les méthodes sont généralement utilisées pour mettre en évidence de possibles structures de cluster lorsqu'aucun *a priori* sur les données n'est disponible. De nombreuses approches existent [Jain et Dubes, 1988; Jain et al., 1999], mais les algorithmes orientés prototype sont certainement les plus populaires. Étant donnés un ensemble de données $X$ et un nombre prédéfini de cluster $c$ choisi par le praticien, de tels algorithmes produisent, en minimisant une fonctionnelle $J$ qui leur est propre, une partition $U$ et un ensemble de $c$ prototypes associés à chaque cluster. Les prototypes les plus simples, lorsque les données sont vectorielles, sont les centres (vecteurs moyenne) des différents clusters. On peut citer, parmi les plus connues de ces méthodes :

- les *K-Moyennes*, aussi appelé *K-Means* ou *Hard C-Means* (HCM, [McQueen, 1967; Jain, 2009]), qui produit des partitions strictes de $\mathbb{M}_{hcn}$,
- les *K-Moyennes Floues*, ou *Fuzzy C-Means* (FCM, [Dunn, 1973; Bezdek, 1981]), qui produit des partitions floues appartenant $\mathbb{M}_{fcn}$,
- les *K-Moyennes Possibilistes*, ou *Possibilistic C-Means* (PCM, [Krishnapuram et Keller, 1993; Davé et Krishnapuram, 1997]), qui produit des partitions possibilistes de $\mathbb{M}_{pcn}$.
- les *K-Moyennes Possibilistes-Floues*, ou *Possibilistic-Fuzzy C-Means* (PFCM, [Pal et al., 2005]), qui produit à la fois une partition possibiliste de $\mathbb{M}_{pcn}$ et une partition floue de $\mathbb{M}_{fcn}$ et qui a pour vocation de corriger certains problèmes inhérents à FCM et PCM.

Les algorithmes FCM et PCM, utilisés pour les expérimentations du Chapitre 4, sont décrits en annexe A.

Tous les algorithmes listés ci-dessus produisent une $c$-partition des données d'entrée, et ce même si ces dernières ne présentent aucune structure en leur sein. Une étape préalable à leur utilisation, souvent négligée par l'utilisateur, consiste donc à vérifier si partitionner les données a du sens ou non, par exemple à l'aide de tests statistiques. Cette problématique, qui n'entre pas dans le cadre de nos travaux, est appelée *cluster*

*tendency* [Jain et Dubes, 1988]. Elle a connu un regain d'intérêt dans la littérature récente, principalement par le biais de méthodes fondées sur le ré-ordonnancement de matrices de similarité entre individus ([Hathaway et Bezdek, 2003; Bezdek et al., 2007; Brouwer, 2009b]).

### 2.3 Cadre d'étude

Dans ce mémoire, on entend par *comparaison de partitions* la mesure du degré d'accord, ou concordance, dont peuvent témoigner deux partitions d'un même ensemble d'individus présentant éventuellement un nombre de clusters différents. Cette problématique, dont l'émergence est liée à l'essor qu'a connu l'informatique depuis la seconde moitié du siècle dernier, prend historiquement sa source dans divers domaines, au sein desquels elle revêt différents aspects et soulève différents enjeux. Ainsi, on en retrouve l'expression :

- en biologie, où l'on citera les travaux de Jaccard [1901], de Sokal et Michener [1958] ou les plus récents de Chavent et al. [2001] dans lesquels une partition d'un ensemble de conifères, établie à partir de la composition chimique de leurs feuilles, est comparée à une partition experte définie par le genre de chaque arbre,
- en psychologie et en sciences sociales, avec notamment les travaux de Brennan et Light [1974] qui proposent d'utiliser une mesure d'accord pour comparer les classements d'une même population d'enfants réalisés par deux psychologues, ou encore ceux de Arabie et Boorman [1973] qui dressent une excellente introduction aux enjeux propre au domaine,
- en développement urbain, avec les travaux de Tavares Pereira et al. [2009] qui tentent d'optimiser, sous diverses contraintes, le découpage d'un territoire donné en un certain nombre de zones ou districts par le biais de mesures de comparaison partitions,
- en management, avec des problématiques telles que la répartition en des classes hétérogènes d'individus formant des classes homogènes, que Charon et al. [2007] proposent de résoudre en minimisant leur mesure de concordance entre partitions,
- et bien évidemment en classification automatique, avec les travaux de Fowlkes et Mallows [1983], de Wallace [1983] et de Rand [1971], ce dernier proposant une méthodologie fondée sur une mesure de comparaison de partitions pour l'étude de classifieurs, ou encore les travaux de Krieger et Green [1999], qui proposent d'utiliser une mesure de comparaison pour le calcul de partition consensus.

À travers la littérature, on distingue ainsi deux types d'utilisations des mesures de comparaison de partitions. La première vise à comparer une partition à une partition de référence, par exemple pour valider le choix d'une méthode de classification supervisée et son paramétrage ou ceux d'un algorithme de clustering. On parle alors de mesures de comparaison *externes*. La seconde tend à comparer deux à deux les éléments d'un ensemble de partitions que l'on souhaite qualifier selon leur similitude ou compatibilité, par exemple pour définir un espace des paramètres pour une méthode de clustering ou comparer des sous-espaces de représentation, et l'on parle alors de mesures de comparaison *relatives*. Bien des mesures relatives peuvent être utilisées comme mesures externes, et certaines mesures externes peuvent de la même manière être utilisées comme mesures relatives. La frontière entre ces deux types de mesures est donc très perméable, et lorsque la distinction n'est pas de mise, nous nous utiliserons le terme plus général de *mesures de comparaison*.

## 2.4 Définitions générales et propriétés des mesures de comparaison de partitions

### 2.4.1 Sur la notion de concordance et de discordance entre deux partitions

Comme nous l'illustrerons plus tard, la notion d'accord entre deux partitions dépend directement de la mesure considérée. Toutefois, l'ensemble de la littérature s'accorde, par exemple dans [Rand, 1971; Arabie et Boorman, 1973], sur la définition des situations dans lesquelles la concordance et la discordance totales sont attendues. Ainsi, on établit habituellement que deux partitions $U$ et $V$ sont en totale concordance, et l'on notera $U \equiv V$, si et seulement si elles sont égales à une permutation des clusters près.

**Définition 6.** *Une matrice de permutation $P_{(\sigma)}$ est une matrice carrée de taille $c$ associée à l'une des permutations $\sigma = \{\sigma(1), \sigma(2), ..., \sigma(c)\}$ de l'ensemble $\{1, 2, ..., c\}$, telle que :*

$$[P_{(\sigma)}]_{ij} = \begin{cases} 1 & si\ i = \sigma(j) \\ 0 & sinon \end{cases} \tag{2.3}$$

*La matrice $P_{(\sigma)}$ est bistochastique*[4] *parce qu'elle ne contient qu'un seul coefficient non-nul égal à 1 par ligne et par colonne. Elle est aussi orthogonale, si bien que $P_{(\sigma)}^{-1} P_{(\sigma)} = {}^{t}P_{(\sigma)} P = I_c$ et que ${}^{t}P_{(\sigma)}$ est la permutation inverse de $P_{(\sigma)}$.*

4. Une matrice bistochastique est une matrice dont les sommes en ligne et les sommes en colonnes sont toutes égales à 1.

Autrement dit, deux partitions sont en concordance totale si $U = P_{(\sigma)}V$, où $P_{(\sigma)}$ est une matrice de permutation. Notons que l'on ne peut avoir $U \equiv V$ que si $U$ et $V$ ont même nombre de clusters. Cependant, il est bien sûr possible qu'une $c-$partition $U' = P_{(\sigma)}\,U$ concorde plus à une $r-$partition $V$ que $U$.

**Exemple 5.** *Soient les deux partitions strictes d'un même ensemble de données :*

$$V_h = \begin{array}{c} \begin{matrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} V_h^1 \\ V_h^2 \\ V_h^3 \end{matrix}$$

*et*

$$V'_h = \begin{array}{c} \begin{matrix} \mathbf{v}'_{h1} & \mathbf{v}'_{h2} & \mathbf{v}'_{h3} & \mathbf{v}'_{h4} & \mathbf{v}'_{h5} \end{matrix} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} V_h'^1 \\ V_h'^2 \\ V_h'^3 \end{matrix}.$$

*On constate que $V_h^1 = V_h'^3$, $V_h^2 = V_h'^1$ et $V_h^3 = V_h'^2$, de sorte que les deux partitions sont équivalentes à une permutation des clusters de l'une d'entre elles près. Cette permutation $\sigma_{V_h,V'_h} = \{3,1,2\}$ se formalise ici très simplement à l'aide de la matrice*

$$P_{(\sigma_{V_h,V'_h})} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

*de telle sorte que $V'_h = P_{(\sigma_{V_h,V'_h})}V_h$, ou encore $V_h = P^{-1}_{(\sigma_{V_h,V'_h})}V'_h = {}^tP_{(\sigma_{V_h,V'_h})}V'_h$.*

Bien évidemment, deux partitions que l'on souhaite comparer ne sont généralement pas équivalentes et la recherche d'une permutation amenant à une égalité totale entre elles se transforme habituellement en la recherche d'une permutation réalisant le meilleur appariement des clusters selon un critère donné. Cet appariement pose toutefois un problème de complexité algorithmique non négligeable. En effet, le nombre de permutations possible d'un ensemble de $c$ clusters est de l'ordre de $c!$, si bien que toute recherche exhaustive devient rapidement impossible. Comme nous le verrons par la suite à la page 18, plusieurs solutions à ce problème ont été proposées.

Un exemple qui illustre bien la notion de désaccord total, que l'on oppose habituellement à celle de concordance totale, est celui des deux partitions $\mathbb{1}_{1n}$ et $I_n$ définies plus tôt à la page 5. Cette notion de discordance reste toutefois à pondérer par la

manière dont peut être abordé, selon les auteurs, le cas de la comparaison de deux partitions présentant un nombre différent de clusters. Dans un tel cas de figure le problème du traitement des $m = \mid c - r \mid$ clusters supplémentaires est soulevé et, bien qu'il soit clair qu'alors les partitions considérées ne peuvent présenter un total accord, il est toutefois probable que la partition arborant le plus grand nombre de clusters constitue un *raffinement* de l'autre (voir Définition 5, page 5). Certains auteurs considèrent alors que ce raffinement témoigne d'une certaine concordance entre les deux partitions et s'attachent à le mesurer, si bien qu'à la lumière unique d'une telle mesure les partitions extrêmes $I_n$ et $\mathbb{1}_{1n}$ décrites plus tôt conservent un certain degré d'accord, voire un degré d'accord certain.

Enfin, on distingue dans la littérature deux types d'approches qui influent directement sur la notion de concordance selon qu'elles orientent les mesures sur les colonnes ou les lignes des matrices de partition. Les mesures du premier type, dont les propriétés seront discutées plus en détails à la page 18, reposent ainsi sur l'idée qu'une partition décrit avant tout une structure relationnelle entre les individus de l'ensemble de données qu'elle partitionne, et que ce sont ces relations qui doivent être privilégiées pour la comparaison. L'un des avantages immédiat est de s'affranchir naturellement du problème de l'appariement des clusters discuté ci-dessus. Au contraire, les approches du second type, décrites à la page 32, privilégient avant tout les clusters. Les mesures qui s'inscrivent dans cette catégorie se focalisent alors sur les similitudes que présentent, deux à deux, les clusters de chaque partition. Dans ce cas, un appariement de ces derniers est généralement nécessaire, et les clusters sont considérés de manière indépendante.

### 2.4.2 Propriétés théoriques et pratiques

Toute fonction visant à évaluer l'accord ou le désaccord entre deux partitions doit permettre de définir formellement une mesure de comparaison. Pour le désaccord, une telle mesure doit alors prendre sa valeur minimale quand les deux partitions sont en accord total et doit croître à mesure de leur discordance, en d'autres termes à mesure que les partitions diffèrent ou s'éloignent l'une de l'autre. Il est alors usuel de définir des *mesures de distance*, ou plus simplement des *distances*, entre matrices de partition $D(U, V) : \mathbb{M}_{pcn} \times \mathbb{M}_{pcn} \to [0, 1]$, que l'on étudie à la lumière des propriétés qu'une métrique doit satisfaire. Lorsque $D$ sera à valeurs dans un autre intervalle fini, on la ramènera dans l'intervalle unité, de sorte que tout ce qui suit ne perde pas en généralité.

**Définition 7.** *[Arabie et Boorman, 1973] Une métrique sur un espace $E$ est une fonction $d : E \times E \to [0, +\infty[$ telle que : pour tous $U, V, W \in E$*

(I-1) $d(U, U) = 0$ (réflexivité),

(I-2) $d(U, V) = 0 \Leftrightarrow U \equiv V$ (identité des indiscernables),

(I-3) $d(U, V) = d(V, U)$ (symétrie),

(I-4) $d(U, W) \leq d(U, V) + d(V, W)$ (subadditivité).

*Si $d$ perd la propriété (I-2) (respectivement l'inégalité triangulaire (I-4)), on dit que $d$ est une* pseudométrique *(respectivement une* semimétrique*). Une métrique $d$ est une* ultramétrique *si elle satisfait l'inégalité suivante, plus forte que (I-4) :*

(I-4′) $d(U, W) \leq \max_V(d(U, V), d(V, W))$ (inégalité ultratriangulaire) .

Lorsqu'a contrario une mesure est construite de sorte de prendre sa valeur maximale pour une concordance totale entre les deux partitions et de décroître avec leur désaccord, on parle alors d'*indice de comparaison* $I(U, V) : \mathbb{M}_{pcn} \times \mathbb{M}_{pcn} \to [0, 1]$, que l'on pourra qualifier selon les propriétés qu'une similarité doit satisfaire. Certains indices de la littérature prennent leurs valeurs dans d'autres intervalles, mais ils peuvent être ramenés à [0,1] de sorte que nos propos s'entendent sans perte de généralité.

**Définition 8.** *[Zadeh, 1971] Une similarité sur un espace $E$ est une fonction $s : E \times E \to [0, 1]$ telle que : pour tous $U, V, W \in E$*

(II-1) $s(U, U) = 1$ (réflexivité) ,

(II-2) $s(U, V) = s(V, U)$ (symétrie) ,

(II-3) $s(U, U) \geq s(U, V)$ (maximalité).

Il découle de la Définition 8 que toute fonction $D_S(U, V) = 1 - S(U, V)$ dérivée d'une similarité $S$ satisfait naturellement les propriétés (I-1) et (I-3). Comme (II-3) est moins forte que (I-4), on qualifiera de préférence tout indice de comparaison $I(U, V)$ selon les propriétés (I-1), (I-2), (I-3) et (I-4) satisfaites par sa mesure de distance duale, comme proposé par Rand [1971].

**Définition 9.** *La mesure de distance duale d'un indice $I$ de comparaison de deux partitions est la fonction définie par $D_I(U, V) = 1 - I(U, V)$. De la même manière, la fonction définie par $I_D(U, V) = 1 - D(U, V)$ est l'indice dual de toute mesure de distance entre deux partitions $D(U, V)$ bornée sur $[0, 1]$.*

Certains auteurs comme Arabie et Boorman [1973] ou Hüllermeier et Rifqi [2009] proposent ainsi que toute mesure de distance $D^\star$, qu'il s'agisse d'une mesure à part entière ou du dual d'un indice de comparaison $I^\star$, doit s'attacher à satisfaire l'intégralité des propriétés d'une métrique. Or, bien qu'il soit clairement désirable que $D^\star$ satisfasse les propriétés de réflexivité *(I-1)* et d'identité des indiscernables *(I-2)*, nous pensons que tel n'est pas tant le cas pour les deux suivantes que sont la symétrie *(I-3)* et la subadditivité *(I-4)*, selon l'usage auquel la mesure considérée est destinée. En effet, alors que la propriété de symétrie est fortement requise pour toute mesure relative dont le but est de comparer deux partitions ordinaires, *i.e.* sur lesquelles aucun *a priori* n'est connu, certains auteurs comme Wallace [1983] et Chavent et al. [2001] ont montré qu'il pouvait être intéressant de considérer des indices asymétriques[5] pour la comparaison d'une partition à une partition de référence, lorsque la mesure est donc vouée à être utilisée comme mesure externe, voir la discussion plus détaillée à la page 23. Quant à la subadditivité de $D^\star$ qui implique le cas échéant que l'indice dual $I^\star$ vérifie que si deux partitions $U$ et $V$ sont proches d'une troisième partition $W$, alors $U$ et $V$ sont elles-mêmes proches, nous pensons qu'elle n'est vraiment nécessaire que dans le cas où la mesure est destinée à être utilisée comme mesure relative pour la comparaison croisée des éléments d'un ensemble de partitions. Elle est ainsi requise pour la méthodologie de représentation par *Positionnement Multidimmensionnel (Multidimensionnal Scaling, (MDS))* proposée par Arabie et Boorman [1973]. Elle reste en revanche superflue dans le cas d'une mesure externe vouée à n'être uniquement utilisée que pour statuer de la proximité d'une partition à une partition de référence.

Nous proposons de considérer une propriété supplémentaire, non formelle, qui devrait être cherchée pour toute distance entre deux partitions ou pour tout indice de comparaison de deux partitions :

(*I-5*). $D^\star(U,V) \ll D^\star(U,W)$ *si* $V$ *est connue pour être bien plus proche de* $U$ *que de* $W$ (*dynamique*).

(*II-4*). $I^\star(U,V) \gg I^\star(U,W)$.

En effet, la connaissance d'une "bonne" dynamique, identifiée empiriquement, permet d'assurer au praticien que l'indice considéré est connu pour présenter de plus ou moins grandes différences de valeur entre la comparaison de partitions proches et celle de partitions éloignées. Ainsi toute décision concernant la compatibilité entre

5. On remarquera par ailleurs que si la propriété de symétrie est absolument requise, il est possible d'obtenir un indice symétrique $I_s$ à partir de n'importe quel indice asymétrique $I$, via : $I_s(U,V) = \mathcal{A}(I(U,V), I(V,U))$, où $\mathcal{A}$ est une opérateur d'agrégation adéquat, par exemple la moyenne arithmétique ou la moyenne géométrique.

deux partitions qu'il aurait à prendre, par exemple en seuillant la valeur de la mesure, serait facilitée.

## 2.5 Une revue des mesures strictes

Bien que le sujet d'étude principal de ce mémoire soit la comparaison de partitions non-strictes d'un même ensemble d'individus, il est nécessaire, ainsi qu'intéressant, de dresser un état de l'art succint mais complet des mesures strictes les plus populaires de la littérature. En effet, c'est sur les bases de nombre de ces dernières que reposent les principales propositions pour les partitions non strictes qui seront revues au Chapitre 3. Nous avons choisi une taxonomie qui nous est propre, fondée sur la distinction faite à la page 13 entre les deux types d'approches (orientées individus et orientées clusters), pour lesquelles la notion même de concordance diverge. Bien évidemment, d'autres taxonomies sont envisageables. Chaque proposition est illustrée par un exemple sur tout ou partie des trois partitions strictes suivantes :

$$U_h = \begin{array}{c} \begin{array}{ccccc} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_4 & \mathbf{u}_5 \end{array} \\ \left( \begin{array}{ccccc} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right) \end{array} \begin{array}{c} U_h^1 \\ U_h^2 \end{array}$$

$$V_h = \begin{array}{c} \begin{array}{ccccc} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 \end{array} \\ \left( \begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right) \end{array} \begin{array}{c} V_h^1 \\ V_h^2 \\ V_h^3 \end{array}$$

$$W_h = \begin{array}{c} \begin{array}{ccccc} \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \mathbf{w}_4 & \mathbf{w}_5 \end{array} \\ \left( \begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array} \begin{array}{c} W_h^1 \\ W_h^2 \\ W_h^3 \\ W_h^4 \end{array}$$

Ces partitions ont été conçues de sorte d'être des raffinements stricts les unes des autres ($W_h \subset V_h \subset U_h$, au sens de la Définition 5) d'un ensemble de données tel que celui représenté à la Figure 2.3.

### 2.5.1 Approches orientées individus

#### 2.5.1.1 Généralités

Une première approche pour la construction de mesures d'accord entre deux partitions repose sur l'observation des appariements entre individus réalisés par chacune

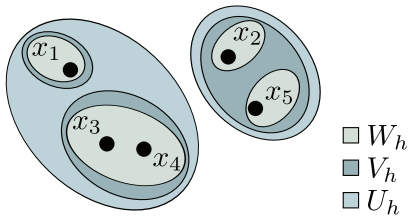



FIGURE 2.3 – Partitions $U_h$, $V_h$ et $W_h$ dans l'espace de représentation des données

des partitions à comparer. C'est pourquoi nous les qualifions de mesures *orientées individus*. Formellement, les mesures de comparaison de deux partitions $U$ et $V$ ainsi dérivées sont fondées sur la construction d'une table (ou matrice) de *contingence-paires* $M(U, V)$ croisant $U$ et $V$ telle que présentée par la Table 2.1.

**Définition 10.** *Un ensemble composé d'exactement deux éléments distincts a et* b *est appelé une* paire. *On le notera entre accolades pour le distinguer du* couple, *composé de deux éléments non forcément distincts dans un ordre déterminé et noté entre parenthèses.*
*Ainsi, deux éléments a et* b *forment une paire* $\{a, b\}$ *unique et peuvent composer jusqu'à quatre couples différents* $(a, a)$, $(a, b)$, $(b, a)$ *et* $(b, b)$.

On peut former $q = \frac{n(n-1)}{2}$ paires d'éléments de $X$. Les termes $m_{11}$, $m_{10}$, $m_{01}$ et $m_{00}$ de la matrice $M(U, V)$ dénombrent les paires $\{\mathbf{x}_k, \mathbf{x}_l\}$ selon l'appartenance conjointe ou non des individus qui la composent aux différents clusters de $U$ et $V$, et sont liés par la relation : $m_{11} + m_{10} + m_{01} + m_{00} = q$. Deux individus distincts $\mathbf{x}_k$ et $\mathbf{x}_l$ de $X$ peuvent :

- appartenir conjointement au même cluster dans $U$ et dans $V$ (accord positif), contribuant à $m_{11}$,
- appartenir au même cluster dans $U$ mais à des clusters différents dans $V$ (désaccord), contribuant à $m_{10}$,
- appartenir à des clusters différents dans $U$ mais au même cluster dans $V$ (désaccord), contribuant à $m_{01}$,
- n'appartenir ni au même cluster dans $U$ ni au même cluster dans $V$ (accord négatif ou compte neutre), contribuant à $m_{00}$.

**Exemple 6.** *Comme on peut facilement l'observer à la Figure 2.3, la partition* $U_h$ *regroupe dans le cluster* $U_h^1$ *les individus* $\mathbf{x}_1, x_3$ *et* $\mathbf{x}_4$ *et dans le cluster* $U_h^2$ *les individus* $\mathbf{x}_2$ *et* $\mathbf{x}_5$, *si bien que l'on en déduit l'ensemble* $U_1 = \{\{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_1, \mathbf{x}_4\}, \{\mathbf{x}_3, \mathbf{x}_4\}, \{\mathbf{x}_2, \mathbf{x}_5\}\}$ *des paires d'individus appartenant au même cluster de* $U$. *De la même*

TABLE 2.1 – Table de contingence-paires $M(U,V)$ de deux partitions $U$ et $V$

| partition | | $V$ | |
|---|---|---|---|
| | # de paires dans | le même cluster | différents clusters |
| $U$ | le même cluster | $m_{11}$ | $m_{10}$ |
| | différents clusters | $m_{01}$ | $m_{00}$ |

*manière on peut construire* $V_1 = \{\{\mathbf{x}_3, \mathbf{x}_4\}, \{\mathbf{x}_2, \mathbf{x}_5\}\}$ *l'ensemble des paires d'individus appartenant au même cluster dans V. Alors, on a :*

$m_{11} =\mid U_1 \bigcap V_1 \mid = card(\{\{\mathbf{x}_3, \mathbf{x}_4\}, \{\mathbf{x}_2, \mathbf{x}_5\}\}) = 2$,

$m_{10} =\mid U_1 \ominus V_1 \mid = card(\{\{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_1, \mathbf{x}_4\}\}) = 2$, *et*

$m_{01} =\mid V_1 \ominus U_1 \mid = card(\emptyset) = 0$.

*En construisant d'une manière analogue les ensembles* $U_0$ *et* $V_0$ *des paires d'individus n'appartenant respectivement pas au même cluster dans U et dans V, on a aussi :*

$m_{00} =\mid U_0 \cap V_0 \mid = card(\{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_5\}, \{\mathbf{x}_3, \mathbf{x}_2\}, \{\mathbf{x}_3, \mathbf{x}_5\}, \{\mathbf{x}_4, \mathbf{x}_2\}, \{\mathbf{x}_4, \mathbf{x}_5\}\}) = 6$,

*si bien que la matrice de contingence-paires vaut* $M(U_h, V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$.

*Évidemment, on a :* $m_{11} + m_{10} + m_{01} + m_{00} = 10 = \frac{5(5-1)}{2} = q$.

L'avantage notoire d'une telle approche réside dans le fait qu'aucun appariement des clusters n'est nécessaire puisque les clusters ne sont pas à proprement parler considérés. Pour la même raison, le problème de la comparaison de partitions présentant un nombre de clusters différents n'a pas à être posé.

Par ailleurs, on sait que dans le cas des partitions strictes, une partition $U$ induit une topologie dite *topologie de partition*, dotant l'espace $X$ qu'elle décrit d'une pseudométrique $d_U : X \times X \rightarrow \{0,1\}$ définie par :

$$d_U(\mathbf{x}_k, \mathbf{x}_l) = \begin{cases} 0 & \text{si } \mathbf{x}_k \text{ et } \mathbf{x}_l \text{ appartiennent au même cluster de U} \\ 1 & \text{sinon} \end{cases} \tag{2.4}$$

si bien que

$$U_1 = \sum_{k=1}^{n} \sum_{l=2}^{k-1} 1 - d(\mathbf{x}_k, \mathbf{x}_l). \tag{2.5}$$

Selon cette acception, les mesures fondées sur les mesures d'appariement entre individus peuvent être perçues comme des mesures de comparaison des espaces pseudométriques $(X, d_U)$ et $(X, d_V)$ décrits par chaque partition $U$ et $V$.

Notons enfin que certains auteurs [Ceccarelli et Maratea, 2008] ont fait des propositions de mesures fondées sur le dénombrement des couples $(\mathbf{x}_k, \mathbf{x}_l)$ plutôt que des paires $\{\mathbf{x}_k, \mathbf{x}_l\}$ d'individus. Les *contingences-couples* correspondantes $m'_{11}$, $m'_{10}$, $m'_{01}$ et $m'_{00}$ intègrent évidemment les contingences-paires si bien que l'on a $m'_{11} + m'_{10} + m'_{01} + m'_{00} = n^2$, mais aussi :

$$m'_{11} = 2m_{11} + n, \tag{2.6}$$

$$m'_{10} = 2m_{10}, \tag{2.7}$$

$$m'_{01} = 2m_{01}, \tag{2.8}$$

$$m'_{00} = 2m_{00}. \tag{2.9}$$

Bien entendu, ce choix est loin d'être optimal d'un point de vue calculatoire et nous pensons par ailleurs que la prise en compte de la contribution à $m_{11}$ des couples $(\mathbf{x}_k, \mathbf{x}_k)$, donc de l'appariement d'un individu avec lui même constitue un biais pour la mesure d'accord, si bien que l'on ne considérera que peu cette approche par la suite.

#### 2.5.1.2 Quelques indices dérivés

À partir des termes de la table des contingence-paires $M(U, V)$ se calculent une multitude de mesures de comparaison, généralement des indices. La Table 2.3 recense les principaux, leur domaine de variation, ainsi que le caractère métrique de la distance complémentaire au sens de la Définition 9. Nous en détaillons ci-après quelques-uns, choisis pour leurs particularités, et les illustrons au travers de la comparaison des partitions strictes $U_h$, $V_h$ et $W_h$ donnés à la page 18.

##### 2.5.1.2.a Indices symétriques

Parmi les indices de comparaison de partition strictes, orientés individus, les plus notables sont assurément l'indice de Rand :

$$RI(U, V) = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} \tag{2.10}$$

et l'indice de Jaccard :

$$JI(U, V) = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}. \tag{2.11}$$

Les distances complémentaires $D_{RI}$ et $D_{JI}$ dérivées de ces deux indices sont des métriques [Rand, 1971; Lipkus, 1999].

**Exemple 7.** *Considérons les partitions $U_h$ et $V_h$ (page 18). À partir de leur matrice de contingence-paires $M(U_h, V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$ donnée à l'Exemple 6, les indices de*

*Rand et de Jaccard correspondants valent :*

$$RI(U_h, V_h) = \frac{2+6}{2+2+0+6} = \frac{4}{5} = 0.800$$

*et*

$$JI(U_h, V_h) = \frac{2}{2+2+0} = \frac{1}{2} = 0.500.$$

La différence fondamentale entre ces deux indices réside dans la prise en compte (par $RI$) ou non (par $JI$) du terme $m_{00}$ pour leur calcul, qui amène à une intéressante discussion au sujet de l'interprétation à donner à ce terme. Comme Wallace [1983] le propose, le nombre $m_{00}$ de paires non jointes dans chaque partition ne témoigne pas avec évidence d'un accord entre les deux partitions. En effet, cette interprétation repose clairement sur l'*a priori* selon lequel la non-appartenance de deux individus au même cluster découle de l'existence d'une quelconque forme de répulsion entre ces deux individus. Or, une telle affirmation dépend directement de la manière dont les partitions sont produites, si bien qu'une certaine réserve est bien de mise lorsque l'on considère des partitions quelconques. De plus, $m_{00}$ tend à croître rapidement avec les nombres $c$ et $r$ de clusters dans chaque partition (c.f. [Wallace, 1983]), amenant les indices qui en tiennent compte à présenter de sérieux problèmes. L'indice de Rand, par exemple, tend ainsi à prendre des valeurs très proches de sa valeur maximale de 1 lorsque le nombre de clusters est important, comme le montrent Fowlkes et Mallows [1983] et Wallace [1983]. Ce comportement a donc mené un certain nombre d'auteurs à considérer $m_{00}$ comme un *compte neutre* qui devrait être délaissé lors de la construction de tout indice de comparaison reposant sur l'appariement d'individus, et à réagir en faveur d'indices tels que celui de Jaccard, dans lequel $m_{00}$ n'intervient pas.

**Exemple 8.** *Si l'on considère aussi la partition $W_h$, on obtient les matrices de contingences-paires $M(U_h, W_h) = \begin{pmatrix} 1 & 3 \\ 0 & 6 \end{pmatrix}$ et $M(V_h, W_h) = \begin{pmatrix} 1 & 1 \\ 0 & 8 \end{pmatrix}$. La Table 2.2 donne l'ensemble des résultats pour toutes les partitions-exemple. Comme discuté plus tôt, on observe bien de fortes variations entre les valeurs respectives des deux indices : l'indice de Jaccard présente une dynamique bien plus importante en indiquant par exemple une faible concordance entre $U_h$ et $W_h$. Par ailleurs, on constate aussi que l'indice de Rand présente une plus grande valeur pour la comparaison de $V_h$ et $W_h$ que celle obtenue pour la comparaison de $V_h$ avec $U_h$. $W_h$ possédant un plus grand nombre de clusters, la tendance de l'indice de Rand à prendre des valeurs très proches de 1 à mesure que le nombre de clusters croît est bien illustrée.*

| $RI$ | $U_h$ | $V_h$ | $W_h$ |
|---|---|---|---|
| $U_h$ | 1 | 0.800 | 0.700 |
| $V_h$ | 0.800 | 1 | 0.900 |
| $W_h$ | 0.700 | 0.900 | 1 |

| $JI$ | $U_h$ | $V_h$ | $W_h$ |
|---|---|---|---|
| $U_h$ | 1 | 0.500 | 0.250 |
| $V_h$ | 0.500 | 1 | 0.500 |
| $W_h$ | 0.250 | 0.500 | 1 |

TABLE 2.2 – $RI$ (à gauche) et $JI$ (à droite) pour les partitions $U_h$, $V_h$ et $W_h$ de la page 18.

#### 2.5.1.2.b Indices asymétriques

Parmi la multitude d'indices dérivés de la table de contingence-paires, une seconde distinction est possible. On constate en effet que plusieurs propositions abandonnent la propriété de symétrie afin de quantifier, plus que l'accord, une quelconque propriété partagée par les deux partitions. L'indice de Rand asymétrique [Chavent et al., 2001], défini par :

$$RAI(U,V) = \frac{m_{11} + m_{00} + m_{10}}{m_{11} + m_{10} + m_{01} + m_{00}} \tag{2.12}$$

a ainsi pour vocation de déterminer à quel point la partition $V$ est plus fine que la partition $U$. Autrement dit, on cherche à mesurer à quel point les classes de $V$ sont incluses dans celles de $U$ ; la prise en compte du nombre de paires $m_{10}$ d'individus disjoints dans $V$ mais réunis dans $U$ dans le calcul de cette inclusion prend ainsi toute sa pertinence. On notera que dans ce cadre, il est clair que $c \leq r$.

**Exemple 9.** *À partir des matrices de contingence-paires des partitions $U_h$, $V_h$ et $W_h$ (page 18), données aux Exemples 8 et 7, on obtient les valeurs d'indice de Rand asymétrique $RAI(U_h, V_h) = RAI(U_h, W_h) = RAI(V_h, W_h) = 1$, qui mettent bien en évidence le raffinement des trois partitions comparées ($W_h \subset V_h \subset U_h$).*

Un autre exemple de mesures asymétriques est celui des deux indices de Wallace [1983], voués à être utilisés comme indices externes. À travers les deux termes $m_{10}$ et $m_{01}$, Wallace souligne l'existence de deux types d'erreurs ($I$ et $II$). Pour une partition $U$ à confronter à une partition de référence $V$, $m_{10}$ peut être considéré comme le nombre d'individus appariés par $U$ qui ne doivent pas l'être eut égard à $V$ : c'est l'erreur de type $I$. L'erreur de type $II$ est définie de la même manière par le le nombre $m_{01}$ d'individus qui ne sont pas appariés dans $U$, alors qu'ils le sont dans $V$. Wallace propose ainsi deux indices quantifiant ces erreurs :

$$BI^I(U,V) = \frac{m_{11}}{m_{11} + m_{10}} \quad \text{et} \quad BI^{II}(U,V) = \frac{m_{11}}{m_{11} + m_{01}}. \tag{2.13}$$

Ces deux indices voient leur valeur augmenter à mesure que diminue respectivement les erreurs de type $I$ et $II$. Hubert et Arabie [1985] proposent une interprétation pro-

babiliste de ces indices. Selon eux, $BI^I$ peut être vu comme la probabilité *a posteriori* que les individus d'une paire tirée au hasard soient dans la même classe dans $V$ sachant qu'ils sont dans la même classe dans $U$ ; $BI^{II}$ est interprété d'une façon analogue.

**Exemple 10.** *À partir des matrices de contingence-paires des partitions $U_h$, $V_h$ et $W_h$ (page 18), données aux Exemples 8 et 7, on obtient $BI^I(U_h, V_h) = 0.500$, qui indique bien que parmi les quatre paires d'individus regroupés deux à deux dans le même cluster par U, seules deux se retrouvent dans V. Pour l'indice $BI^{II}$, on observe que :*

$$BI^{II}(U_h, V_h) = BI^{II}(U_h, W_h) = BI^{II}(V_h, W_h) = 1.$$

*Sachant que les partitions sont conçues de telle sorte que $W_h \subset V_h \subset U_h$, on remarque que l'indice $B^{II}$ prend sa valeur maximale dans ce cas puisque les paires appariées dans le même cluster par les partitions les plus fines le sont a fortiori par les partitions les plus grossières.*

Il est intéressant de noter que les indices de Kulczynski [1927] et de Fowlkes et Mallows [1983] tels que définis dans la Table 2.3 se définissent respectivement comme les moyennes arithmétique et géométrique de ces deux indices et en constitue ainsi un compromis symétrique.

#### 2.5.1.2.c Indices corrigés pour la chance

Parce qu'il est évident qu'une partie des appariements comptabilisés par $m_{11}$ et $m_{00}$ peut n'être due qu'au hasard, certains auteurs [Hubert et Arabie, 1985; Chavent et al., 2001] ont proposé de modifier plusieurs indices de la littérature afin d'en tenir compte. Tout indice peut ainsi être corrigé afin que sa valeur soit nulle lorsque les partitions comparées sont générés aléatoirement, via la formule suivante [Hubert et Arabie, 1985] :

$$I_c(U, V) = \frac{I(U, V) - E(I(U, V))}{1 - E(I(U, V))}, \tag{2.14}$$

où $E(I)$ est l'espérance de l'indice sous l'hypothèse nulle que les termes $n_{ij}$ de la table de contingence sont tirés aléatoirement selon une loi hypergéométrique multivariée de paramètres[6] $(n, n_{\bullet j}, n_{i\bullet})$. On retiendra pour l'exemple l'indice de Rand Ajusté, si souvent cité dans la littérature, définit par [Hubert et Arabie, 1985] :

$$ARI(U, V) = \frac{m_{11} - \frac{(m_{11}+m_{10})(m_{11}+m_{01})}{m_{11}+m_{10}+m_{01}+m_{00}}}{\frac{(m_{11}+m_{10})+(m_{11}+m_{01})}{2} - \frac{(m_{11}+m_{10})(m_{11}+m_{01})}{m_{11}+m_{10}+m_{01}+m_{00}}}, \tag{2.15}$$

6. Effectif total de $X$, des clusters de $V$, de $U$, définis par la Table 2.4, page 29

**Exemple 11.** *Considérons les partitions $U_h$ et $V_h$ (page 18). À partir de leur matrice de contingence-paires $M(U_h, V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$ donnée à l'Exemple 6, la valeur de leur indice de Rand ajusté est $ARI(U_h, V_h) = 0.545$, alors que $RI(U_h, V_h) = 0.800$ (voir Exemple 7). Ceci illustre bien que la correction diminue généralement la valeur de l'indice.*

Notons toutefois que l'hypothèse nulle sur laquelle reposent ces corrections est parfois contestée [Wallace, 1983]. Il a aussi été montré par Albatineh et al. [2006] qu'après correction pour la chance, un certain nombre d'indices connus se confondent pour devenir strictement équivalents, comme l'indice de Kulczynski [1927] et celui de McConnaughey [1964].

#### 2.5.1.3 Aspects calculatoires

Les termes $m_{11}$, $m_{10}$, $m_{01}$ et $m_{00}$ de la matrice de contingence-paires $M(U, V)$ peuvent être calculés de différentes manières, strictement équivalentes, décrites ci-après.

##### 2.5.1.3.a Approche ensembliste

La méthode la plus naturelle pour le calcul de $M(U, V)$ est très certainement celle fondée sur une approche ensembliste qui repose sur la définition des quatre ensembles suivants [Campello, 2007] :

- $U_1$ et $V_1$ : les ensembles regroupant les paires d'individus $(\mathbf{x}_k, \mathbf{x}_l)$ appartenant au même cluster dans $U$ et dans $V$, respectivement,
- $U_0$ et $V_0$ : les ensembles regroupant les paires d'individus $(\mathbf{x}_k, \mathbf{x}_l)$ appartenant à différents clusters dans $U$ et dans $V$, respectivement,

tels que :

$$U_1 = \bigcup_{i=1}^{c} U^{ii} \quad \text{et} \quad V_1 = \bigcup_{i=1}^{r} V^{ii}, \tag{2.16}$$

$$U_0 = \bigcup_{\substack{i \neq j \\ i,j=1}}^{c} U^{ij} \quad \text{et} \quad V_0 = \bigcup_{\substack{i \neq j \\ i,j=1}}^{r} V^{ij}, \tag{2.17}$$

où $U^{ii}$ (respectivement $V^{ii}$) est l'ensemble des paires d'individus appartenant au même $i$-ème cluster de $U$ (respectivement $V$), et $U^{ij}$ (respectivement $V^{ij}$) est l'ensemble des paires appartenant aux $i$-ème et $j$-ème ($j \neq i$) clusters de $U$ (respectivement $V$). Les quatre termes $\{m_{\alpha\beta} : \alpha, \beta \in \{0, 1\}\}$ de $M(U, V)$ sont alors simplement donnés par :

$$m_{\alpha\beta}(U, V) = card(U_\alpha \cap V_\beta). \tag{2.18}$$

TABLE 2.3 – Principaux indices de comparaison orientés individus

| Indice ($I$) | Formule | Intervalle | $D_I = 1 - I$ |
|---|---|---|---|
| | Indices symétriques | | |
| Indice de Rand [1971] ($RI$) | $\frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}}$ | $[0, 1]$ | métrique |
| Indice de Russel et Rao [1940] ($RRI$) | $\frac{m_{11}}{m_{11} + m_{10} + m_{01} + m_{00}}$ | $[0, 1]$ | semimétrique |
| Indices de Gower et Legendre [1986]($GLI$) | $\frac{m_{11} + m_{00}}{m_{11} + \theta(m_{10} + m_{01}) + m_{00}}$ | $[0, 1]$ | métrique si $\theta \geq 1$ |
| Indice de Jaccard [1901] ($JI$) | $\frac{m_{11}}{m_{11} + m_{10} + m_{01}}$ | $[0, 1]$ | métrique |
| Indice de Dice [1945] ($DI$) | $\frac{2m_{11}}{2m_{11} + m_{10} + m_{01}}$ | $[0, 1]$ | semimétrique |
| Indice de Sokal et Sneath [1963] ($SSI_1$) | $\frac{m_{11}}{m_{11} + 2(m_{10} + m_{01})}$ | $[0, 1]$ | métrique |
| Indice de Sokal et Sneath [1963] ($SSI_2$) | $\frac{1}{4}(\frac{m_{11}}{m_{11} + m_{10}} + \frac{m_{11}}{m_{11} + m_{01}} + \frac{m_{00}}{m_{10} + m_{00}} + \frac{m_{00}}{m_{01} + m_{00}})$ | $[0, 1]$ | semimétrique |
| Indice de Fowlkes et Mallows [1983] ($FMI$) | $\frac{m_{11}}{\sqrt{(m_{11} + m_{10})(m_{11} + m_{01})}}$ | $[0, 1]$ | semimétrique |

*suite page suivante*

*suite de la page précédente*

| Indice ($I$) | Formule | Intervalle | $D_I = 1 - I$ |
|---|---|---|---|
| Indice de Kulczynski [1927] ($KI$) | $\frac{1}{2}(\frac{m_{11}}{m_{11}+m_{10}} + \frac{m_{11}}{m_{11}+m_{01}})$ | $[0,1]$ | semimétrique |
| Indice de McConnaughey [1964] ($MCI$) | $\frac{m_{11}^2 - m_{10}m_{01}}{(m_{11}+m_{10})(m_{11}+m_{01})}$ | $[-1,1]$ | $\times$ |
| Indice de Rand Ajusté [Hubert et Arabie, 1985] ($ARI$) | $\frac{m_{11} - \frac{(m_{11}+m_{10})(m_{11}+m_{01})}{m_{11}+m_{10}+m_{01}+m_{00}}}{\frac{(m_{11}+m_{10})+(m_{11}+m_{01})}{2} - \frac{(m_{11}+m_{10})(m_{11}+m_{01})}{m_{11}+m_{10}+m_{01}+m_{00}}}$ | $[-1,1]$ | $\times$ |
| $\Gamma$-statistics [Jain et Dubes, 1988] ($\Gamma$) | $\frac{m_{11}m_{00} - m_{10}m_{01}}{\sqrt{(m_{11}+m_{10})(m_{11}+m_{01})(m_{10}+m_{00})(m_{01}+m_{00})}}$ | $[-1,1]$ | $\times$ |
| Indice de Sokal et Sneath [1963] ($SSI_3$) | $\frac{m_{11}m_{00}}{\sqrt{(m_{11}+m_{10})(m_{11}+m_{01})(m_{10}+m_{00})(m_{01}+m_{00})}}$ | $[-1,1]$ | $\times$ |
| | Indices asymétriques | | |
| Indice de Rand Asymétrique [Chavent et al., 2001] ($RAI$) | $\frac{m_{11}+m_{00}+m_{01}}{m_{11}+m_{10}+m_{01}+m_{00}}$ | $[0,1]$ | $\times$ |
| Indices de Wallace [1983] ($B^I$) et ($B^{II}$) | $\frac{m_{11}}{m_{11}+m_{10}}$<br>$\frac{m_{11}}{m_{11}+m_{01}}$ | $[0,1]$ | $\times$ |

**Exemple 12.** *Considérons les partitions $U_h$ et $V_h$ (page 18). Les quatre ensembles de paires définis par (2.16) et (2.17) sont :*

$U_1 = \{(\mathbf{x}_1, \mathbf{x}_3), (\mathbf{x}_1, \mathbf{x}_4), (\mathbf{x}_3, \mathbf{x}_4), (\mathbf{x}_2, \mathbf{x}_5)\}$,

$V_1 = \{(\mathbf{x}_3, \mathbf{x}_4), (\mathbf{x}_2, \mathbf{x}_5)\}$,

$U_0 = \{(\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1, \mathbf{x}_5), (\mathbf{x}_3, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_5), (\mathbf{x}_4, \mathbf{x}_2), (\mathbf{x}_4, \mathbf{x}_5)\}$, *et*

$V_0 = \{(\mathbf{x}_1, \mathbf{x}_3), (\mathbf{x}_1, \mathbf{x}_4), (\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1, \mathbf{x}_5), (\mathbf{x}_3, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_5), (\mathbf{x}_4, \mathbf{x}_2), (\mathbf{x}_4, \mathbf{x}_5)\}$.

*La matrice de contingence-paires* $M(U_h, V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$ *peut alors être facilement retrouvée par (2.18).*

#### 2.5.1.3.b Approche contigentielle

Ces termes peuvent aussi être calculés à partir de la *matrice de contingence* $N(U, V)$ croisant les clusters des deux partitions, et illustrée par la partie intérieure de la Table 2.4. Chaque terme $n_{ij}$ de $N(U, V)$ représente le nombre d'individus appartenant au $i$-ème cluster de $U$ et au $j$-ème cluster de $V$, soit :

$$n_{ij} = \mid U^i \bigcap V^j \mid . \tag{2.19}$$

Les partitions étant strictes, ce cardinal de l'intersection entre chaque paire $\{U^i, V^j\}$ de clusters s'obtient par produit scalaire des lignes des partitions, de sorte que $N(U, V)$ se définit aisément.

**Définition 11.** *[Brouwer, 2009a] Étant données deux partitions strictes $U$ et $V$, leur* matrice de contingence *est la matrice de taille $c \times r$, définie par :*

$$N(U, V) = U\ {}^tV. \tag{2.20}$$

De cette construction, les termes de la table de contingence-paires $M(U, V)$ s'expriment comme suit :

$$m_{11} = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}(n_{ij} - 1), \tag{2.21}$$

$$m_{10} = \frac{1}{2}\Big(\sum_{j=1}^{c} n_{i\bullet}^2 - \sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}^2\Big), \tag{2.22}$$

$$m_{01} = \frac{1}{2}\Big(\sum_{j=1}^{r} n_{\bullet j}^2 - \sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}^2\Big), \tag{2.23}$$

$$m_{00} = \frac{1}{2}\left((\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij})^2 + \sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}^2 - \Big(\sum_{i=1}^{c} n_{i\bullet}^2 + \sum_{j=1}^{r} n_{\bullet j}^2\Big)\right). \tag{2.24}$$

Remarquons que puisque $\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij} = n$, l'équation (2.24) est souvent présentée par des auteurs de la littérature, par exemple Anderson et al. [2010], sous une forme réduite :

$$m_{00} = \frac{1}{2}\left(n^2 + \sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}^2 - \left(\sum_{i=1}^{c} n_{i\bullet}^2 + \sum_{j=1}^{r} n_{\bullet j}^2\right)\right). \tag{2.25}$$

**Exemple 13.** *Considérons les partitions $U_h$ et $V_h$ (page 18). Leur matrice de contingence vaut $N(U_h, V_h) = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ par (2.20), d'où on déduit leur matrice de contingence-paires $M(U_h, V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$ par (2.21), (2.22), (2.23) et (2.24).*

#### 2.5.1.3.c Approche coïncidentielle

De manière équivalente, la table de contingence-paires peut être calculée à partir des *coïncidences* d'appartenance, dans chacune des deux partitions, des paires d'individus.

**Définition 12.** *La* **matrice de coïncidence** $\Psi_U$ *d'une matrice de partition stricte $U$, est la matrice de taille $(n \times n)$, dont le terme général est tel que :*

$$\psi_{U,kl} = \begin{cases} 1 & \text{si } \mathbf{x}_k \text{ et } \mathbf{x}_l \text{ appartiennent au même cluster de } U \\ 0 & \text{sinon} \end{cases} \tag{2.26}$$

*Généralement, on calcule directement cette matrice à partir de la matrice de partition qu'elle caractérise :*

$$\Psi_U = {}^t U\, U. \tag{2.27}$$

TABLE 2.4 – Table de contingence $N(U, V)$ croisant les clusters de deux partitions $U$ et $V$

| partition | | $V$ | | | | |
|---|---|---|---|---|---|---|
| | clusters | $V^1$ | $V^2$ | $\cdots$ | $V^r$ | *total* |
| $U$ | $U^1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1r}$ | $n_{1\bullet}$ |
| | $U^2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2r}$ | $n_{2\bullet}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $U^c$ | $n_{c1}$ | $n_{c2}$ | $\cdots$ | $n_{cr}$ | $n_{c\bullet}$ |
| | *total* | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet r}$ | $n_{\bullet\bullet} = n$ |

On notera que (2.27) est la négation directe de la pseudométrique induite par $U$ et définie plus tôt par (2.4) page 20 .

À partir des deux matrices de coïncidence $\Psi_U$ et $\Psi_V$ de deux partitions $U$ et $V$ à comparer, les quatre termes $\{m_{\alpha\beta} : \alpha,\beta \in \{0,1\}\}$ de $M(U,V)$ sont donnés par une formule unique :

$$m_{\alpha\beta}(\Psi_U,\Psi_V) = \sum_{k=2}^{n}\sum_{l=1}^{k-1}\Big((1-\alpha)+(2\alpha-1)\ \psi_{U,kl}\Big) \wedge \Big((1-\beta)+(2\beta-1)\ \psi_{V,kl}\Big) \quad (2.28)$$

où $\wedge$ est l'opérateur de conjonction logique (ET booléen).

**Exemple 14.** *Les matrices de coïncidence respectives des deux partitions $U_h$ et $V_h$ décrites page 18 sont, par (2.27) :*

$$\Psi_{U_h} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \text{ et } \Psi_{V_h} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

*Par (2.28), on retrouve leur matrice de contingence-paires* $M(U_h,V_h) = \begin{pmatrix} 2 & 2 \\ 0 & 6 \end{pmatrix}$.

Nous montrons maintenant qu'il est possible de calculer les quatre termes de la matrice de contingence-couples $M'(U,V)$ à partir de produits des matrices de coïncidence $\Psi_U$ et $\Psi_V$. Comme ils sont liés à ceux de la matrice contingence-paires $M(U,V)$ par (2.6)(2.7)(2.8)(2.9), il est possible de calculer les termes de cette dernière de la même manière.

**Théorème 1.** *Soient deux partitions strictes $U$ et $V$, de matrice de coïncidence respective $\Psi_U$ et $\Psi_V$. Les quatre termes $\{m'_{\alpha\beta} : \alpha,\beta \in \{0,1\}\}$ de leur matrice de contingence-couples $M'(U,V)$ sont donnés par :*

$$m'_{11} = trace(\Psi_U\Psi_V) \quad (2.29)$$
$$m'_{10} = trace(\Psi_U\Psi_U) - trace(\Psi_U\Psi_V) \quad (2.30)$$
$$m'_{01} = trace(\Psi_V\Psi_V) - trace(\Psi_U\Psi_V) \quad (2.31)$$
$$m'_{00} = trace((1-\Psi_U)(1-\Psi_V)) \quad (2.32)$$

*et les quatre termes $\{m_{\alpha\beta} : \alpha,\beta \in \{0,1\}\}$ de leur matrice de contingence-couples*

*$M'(U,V)$ sont donnés par :*

$$m_{11} = \frac{trace(\Psi_U\Psi_V) - n}{2} \tag{2.33}$$

$$m_{10} = \frac{trace(\Psi_U\Psi_U) - trace(\Psi_U\Psi_V)}{2} \tag{2.34}$$

$$m_{01} = \frac{trace(\Psi_V\Psi_V) - trace(\Psi_U\Psi_V)}{2} \tag{2.35}$$

$$m_{00} = \frac{trace((1-\Psi_U)(1-\Psi_V))}{2} \tag{2.36}$$

*Preuve de l'équation (2.29).* Pour toutes partitions $U$ et $V \in \mathbb{M}_{hcn}$, on a :

$$\begin{aligned} trace(\Psi_U\Psi_V) &= \sum_{k=1}^{n} < \psi_U^k, \psi_V^k > \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{U,kl}\,\psi_{V,kl} \end{aligned}$$

Par (2.27), les produits $\psi_{U,kl}\,\psi_{V,kl}$ valent 1 si $\mathbf{x}_k$ et $\mathbf{x}_l$ sont dans le même cluster dans $U$ et dans $V$, 0 sinon. Par conséquent, on a bien : $trace(\Psi_U\Psi_V) = m'_{11}$. □

L'équation (2.32) se prouve de manière analogue.

*Preuve de l'équation (2.30).* Pour toute partition $U \in \mathbb{M}_{hcn}$, on a :

$$trace(\Psi_U\Psi_U) = \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{U,kl}^2$$

Par (2.27), les produits $\psi_{U,kl}\,\psi_{U,kl}$ valent 1 si $\mathbf{x}_k$ et $\mathbf{x}_l$ sont dans le même cluster de $U$, 0 sinon, qu'ils soient ou non dans le même cluster de $V$. Par conséquent, on a : $trace(\Psi_U\Psi_U) = m'_{11} + m'_{10}$, et on a bien $m'_{10} = trace(\Psi_U\Psi_U) - trace(\Psi_U\Psi_V)$. □

Une preuve analogue permet d'établir (2.31).

Il est également possible de relier les approches coïncidentielle et contingentielle, sans passer par les contingence-paires ou les contingence-couples. Les formules de passage ont été établies [Kendall et Stuart, 1961; Marcotorchino, 1984] :

$$\sum_{i=1}^{c} n_{i\bullet}^2 = \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{U,kl} \tag{2.37}$$

$$\sum_{j=1}^{r} n_{\bullet j}^2 = \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{V,kl} \tag{2.38}$$

$$\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}^2 = \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{U,kl}\ \psi_{V,kl}. \tag{2.39}$$

### 2.5.2 Approches orientées clusters

Comme évoqué plus tôt (page 13), les approches orientées clusters s'attachent à mesurer la concordance entre deux partitions en se focalisant avant tout sur les similitudes que peuvent présenter, deux à deux, leurs clusters respectifs. Contrairement aux mesures orientées individus que nous venons de présenter, ces approches sont confrontées à deux problèmes majeurs :

- la recherche à moindre coût du meilleur appariement (ou de la meilleure permutation) entre les clusters de chaque partition,
- la comparaison de partitions présentant un nombre de clusters différents.

#### 2.5.2.1 Mesures fondées sur une mesure de compatibilité entre clusters

Les mesures de comparaison de partitions décrites ici sont fondées sur l'observation d'une certaine correspondance entre les clusters des deux partitions comparées. Ces mesures de la littérature ne sont finalement qu'une manière particulière d'agréger les valeurs de *mesures de compatibilité* entre clusters.

Dans la suite, on notera $\mathcal{C}(X)$, l'ensemble des sous-ensembles définis sur un référentiel $X$, et on se rappellera que dans le cadre de nos travaux ce référentiel est un ensemble $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ à partir desquelles sont définies des partitions.

**Définition 13.** *Tversky [1977] Une* mesure de compatibilité *entre deux sous-ensembles $A$ et $B$ de $\mathcal{C}(X)$ est une fonction $M(A,B) = F(a,b,c)$ avec $a = f(A \cap B)$, $b = f(A \ominus B)$ et $c = f(B \ominus A)$, où $f$ est une mesure additive*[7] [8]*. On distingue deux modèles de construction pour $F$ :*

- *le* modèle contraste *: $F_C(a,b,c) = \theta a - \alpha b - \beta c$, avec $\theta, \alpha, \beta \geq 0$,*
- *le* modèle rapport *: $F_R(a,b,c) = \frac{a}{a - \alpha b - \beta c}$, avec $\alpha, \beta \geq 0$.*

*On notera $M_C(U,V)$ et $M_R(U,V)$ les mesures construites à partir des modèles contraste et rapport, respectivement.*

##### 2.5.2.1.a Mesures fondées sur l'erreur de classification

Ces mesures reposent sur la quantification de l'erreur de classification réalisée par une partition au regard d'une seconde partition de référence. Parmi celles-ci, l'indice proposé par Meilă et Heckerman [2001] est défini par :

$$\mathcal{HC}(U,V) = \begin{cases} 1 - \frac{1}{n} \underset{\sigma \in \Sigma}{max} \sum_i^c \mid U^i \cap (P_{(\sigma)}V)^j \mid, & \text{si } c \leq r \\ \mathcal{HC}(V,U) & \text{sinon} \end{cases} \tag{2.40}$$

7. Une mesure $f$ est dite additive si $f(A \cup B) = f(A) + f(B)$ lorsque $A$ et $B$ sont disjoints.
8. Dans ce cadre, $f$ une généralement fonction cardinale.

où $\Sigma$ est l'ensemble des permutations $\sigma$ possibles sur $V$. Notons encore l'indice asymétrique de Larsen et Aone [1999] :

$$\mathcal{L}(U,V) = \frac{1}{c}\sum_{i}^{c} \underset{j\in[1,2,\dots,r]}{max} \frac{2 \mid U^i \cap V^j \mid}{2 \mid U^i \cap V^j \mid - \mid U^i - V^j \mid - \mid V^i - U^j \mid}, \tag{2.41}$$

ou la métrique de Dongen [2000] :

$$\mathcal{D}(U,V) = 2n - \sum_{i}^{c} \underset{j\in[1,2,\dots,r]}{max} \mid U^i \cap V^j \mid - \sum_{j}^{r} \underset{i\in[1,2,\dots,c]}{max} \mid U^i \cap V^j \mid, \tag{2.42}$$

Cette dernière distance est bornée par $2n$, si bien qu'il est aisé de ramener ses valeurs à l'intervalle unitaire.

Comme nous le montrons à la Table 2.5 (page 35), chacune de ces propositions peut être exprimée comme l'agrégation d'une mesure de compatibilité $M$ se conformant à la Définition 13. Toutes les mesures de cette famille ont été critiquées du fait que, ne tenant compte uniquement que de la contribution des individus correctement classés, elles ignorent complètement les autres, ne considérant ainsi aucunement tout possible accord entre les deux partitions à propos de ces individus, comme l'illustre l'Exemple 15 ci-dessous issu de [Meilă, 2007].

**Exemple 15.** *Considérons les trois 3-partitions de la Figure 2.4. Il est raisonnable d'affirmer que V" est plus proche de U" que ne l'est W" puisqu'on retrouve entre U" et V" des individus qui, bien que classés dans des clusters différents d'une partition à l'autre restent regroupés au sein d'un même cluster dans V" alors qu'ils ne le sont pas dans W". Or le calcul de la mesure $\mathcal{H}$ nous amène à constater que :*

$$\mathcal{H}(U'',V'') = \mathcal{H}(U'',W'') = \tfrac{24}{36}$$

*Comme évoqué plus tôt, ce comportement provient du fait que l'indice $\mathcal{H}$ ne considère que les individus bien classés et non les relations entre les individus eux-même comme pourrait le faire une mesure d'appariement telles que celles décrites en 2.5.1. Ce constat est valable pour les mesures $\mathcal{L}$ et $\mathcal{D}$.*

#### 2.5.2.1.b Mesures fondées sur la distance d'édition

La famille des mesures fondées sur la distance d'édition repose sur l'idée simple du comptage du nombre minimal d'opérations d'édition à appliquer à l'une des partitions à comparer pour la transformer en une seconde partition. Dans ce cadre, sont alors uniquement permises les opérations d'ajout et de retrait d'individus, ainsi que

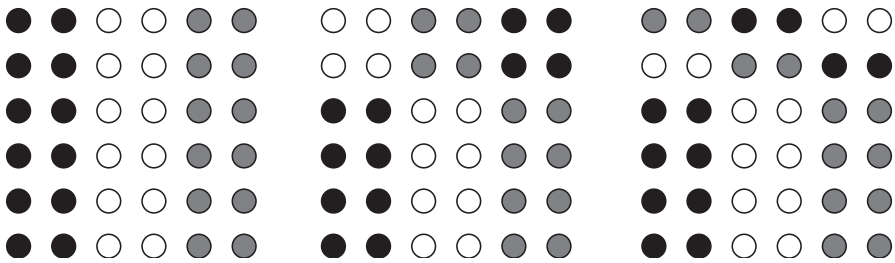


FIGURE 2.4 – Trois partitions ($U''$, $V''$, $W''$) en 3 clusters (●,●,○) d'un même ensemble de 36 individus.

celles de fusion et de fission des clusters de $U$ [Charon et al., 2006]. De nombreuses distances peuvent être construites à partir de ces transformations élémentaires, selon que tout ou partie de ces dernières sont considérées [Day, 1981]. Ainsi, la *distance de transfert* $TD(U,V)$, introduite par Régnier [1965] puis décrite par Denoeud et Guenoche [2006] et certainement la plus populaire de ces mesures, se restreint aux simples opérations de retrait et d'ajout d'individus pour mesurer la similitude entre deux partitions $U$ et $V$. Celles-ci permettent de quantifier le nombre d'éléments à transférer d'un cluster de $U$ vers un cluster de $V$, potentiellement vide[9], pour transformer $U$ en $V$, ou inversement.

À partir de la σ-concordance définie par :

$$c_\sigma(U,V) = \sum_{i}^{c} \mid U^i \cap (P_{(\sigma)}V)^j \mid \tag{2.43}$$

où $P_{(\sigma)}$ est une matrice de permutation (Définition 6, page 13) la distance de transfert, qui est une métrique, s'exprime :

$$TD(U,V) = n \; - \; \max_{\sigma\in\Sigma} \; c_\sigma(U,V) \tag{2.44}$$

où $\Sigma$ est l'ensemble des permutations possibles et $\max_{\sigma\in\Sigma} c_\sigma(U,V)$ est la concordance entre les partitions $U$ et $V$, caractérisant le nombre d'éléments conjoints aux clusters de $U$ et de $V$ selon la permutation optimale $P_{(\sigma)}$ et n'ayant donc pas à être déplacés pour transformer $U$ en $V$. Lorsque $U$ et $V$ ne présentent pas le même nombre de clusters ($c \neq r$), les auteurs proposent d'injecter $m = max(c,r) - min(c,r)$ clusters vides dans la partition en présentant le moins afin de réaliser un appariement total. On remarquera que bien que la définition de la σ-concordance (2.43) est sensiblement analogue à celle de l'indice $\mathcal{H}$ de Meilă et Heckerman [2001] défini par l'équation (2.40), ces deux mesures divergent par l'appariement qu'elles réalisent.

9. Afin de gérer le cas où les partitions ont un nombre de clusters différent.

Celui-ci est ainsi total pour la $\sigma$-concordance alors qu'il n'est que partiel pour l'indice de Meilă et Heckerman [2001]. La distance de transfert (2.44) peut ainsi être vue comme une mesure de l'erreur de classification présentant une meilleure granularité que les méthodes précédentes, palliant ainsi leur défaut, et peut donc se reformuler à partir de mesures de compatibilité.

TABLE 2.5 – Reformulation de $\mathcal{H}$, $\mathcal{L}$, $\mathcal{D}$ et $TD$ à l'aide de mesures de compatibilité

| Mesure | $\theta$ | $\alpha$ | $\beta$ |
|---|---|---|---|
| $\mathcal{H}(U,V) = 1 - \frac{1}{n} \max_{\sigma\in\Sigma} \sum_i^c M_C(U^i, V^j)$ | 1 | 0 | 0 |
| $\mathcal{L}(U,V) = \frac{1}{c} \sum_i^c \max_{j\in[1,2,\dots,r]} M_R(U^i, V^j)$ | $\times$ | $\frac{1}{2}$ | $\frac{1}{2}$ |
| $\mathcal{D}(U,V) = 2n - \sum_i^c \max_{j\in[1,2,\dots,r]} M_C(U^i, V^j) - \sum_j^r \max_{i\in[1,2,\dots,c]} M_C(U^i, V^j)$ | 1 | 0 | 0 |
| $TD(U,V) = n - \max_{\sigma\in\Sigma} \sum_i^c M_C(U^i, V^j)$ | 1 | 0 | 0 |

**Exemple 16.** *Considérons les partitions $U_h$ et $V_h$ (page 18), rappelées ici :*

$$U_h = \begin{array}{c} \begin{matrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_4 & \mathbf{u}_5 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} U_h^1 \\ U_h^2 \end{matrix} \qquad V_h = \begin{array}{c} \begin{matrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} V_h^1 \\ V_h^2 \\ V_h^3 \end{matrix}$$

*Puisque $U_h$ présente moins de clusters que $V_h$, nous l'agrémentons de $m = r - c = 3 - 2 = 1$ cluster vide amenant à la partition dégénérée $\widetilde{U_h}$ :*

$$\widetilde{U_h} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} \widetilde{U_h^1} \\ \widetilde{U_h^2} \\ \widetilde{U_h^3} \end{matrix}$$

*Intuitivement, on établit que la meilleure permutation $\sigma_1$ entre les clusters de $\widetilde{U_h}$ et de $V_h$ est alors celle appariant $\widetilde{U_h^1}$, $\widetilde{U_h^2}$ et $\widetilde{U_h^3}$ avec $V_h^2$, $V_h^3$ et $V_h^1$, respectivement, et pour lequel $c_{\sigma_1} = 4$. Elle est décrite par la matrice de permutation $P_{(\sigma_1)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$, si bien que l'on a : $V'_h = P_{(\sigma_1)} V_h = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$.*

*Pour transformer $\widetilde{U_h}$ en $V'_h$, il suffit alors de :*

- *retirer l'individu $x_1$ de $\widetilde{U_h^1}$,*
- *d'ajouter ce même individu à $\widetilde{U_h^3}$,*

*soit donc de transférer $x_1$ de $\widetilde{U_h^1}$ à $\widetilde{U_h^2}$.*

*Puisqu'un seul transfert est requis, on a finalement : $TD(U_h, V_h) = 1 = n - c_{\sigma_1}$.*

La recherche de la $\sigma$-concordance maximale, et donc de la meilleure permutation $\sigma$, est bien évidemment une tâche coûteuse d'un point de vue calculatoire si elle est exhaustive. Ainsi, Day [1981] préconise de modéliser le problème selon la théorie des graphes en construisant le *graphe biparti complet* dont l'ensemble des clusters de $U$ et $V$ forment les deux bipartitions.

**Définition 14.** *Un graphe est dit* biparti complet *s'il existe une partition de son ensemble de sommets en deux sous-ensembles $U$ et $V$ telle que chaque sommet de $U$ est relié à chaque sommet de $V$.*

En valuant les arêtes de ce graphe par le cardinal de l'intersection de chaque paire de clusters $\{U^i, V^j\}$ alors le problème de la recherche du nombre de transferts minimum à effectuer pour transformer $U$ en $V$ revient à celui de la recherche du couplage[10] de poids maximum dans le graphe construit et peut alors être résolu bien plus rapidement par des algorithmes tels que la méthode Hongroise [Kuhn, 1955]. La Figure 2.5 illustre cette construction pour les partitions de l'Exemple 16.

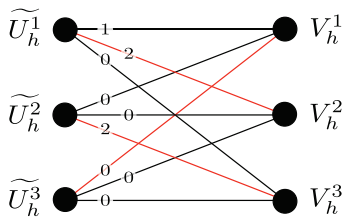



FIGURE 2.5 – Couplage de poids maximum (en rouge) du graphe biparti complet construit à partir des partitions $\widetilde{U_h}$ et $V_h$. Les arcs sont valués par le cardinal de l'intersection de chaque paire de clusters.

Enfin, il a été montré qu'étant donnée deux partitions $U$ et $V$ quelconques à respectivement $c$ et $r$ clusters, il existe une concordance minimum $c_{min}(U, V)$ bornant ainsi supérieurement $TD(U, V)$ quelque soit $V$. Cette concordance minimale est donnée par :

$$c_{min}(U,V) = \begin{cases} c + r - n & \text{si } n \leq c + r - 2 \\ \lceil \frac{n + r - c}{r} \rceil & \text{si } c + r - 1 \leq n \leq (c-1)r \\ \lceil \frac{n}{r} \rceil & \text{si } (c-1)r < n \end{cases} \tag{2.45}$$

où $\lceil . \rceil$ représente la partie entière par excès, et où $c \leq r$.

10. Un couplage d'un graphe est un ensemble d'arêtes de ce graphe qui n'ont pas de sommets en commun. Il s'agit donc bien ici d'un appariement des clusters.

#### 2.5.2.2 Mesures fondées sur la théorie de l'information

##### 2.5.2.2.a Information mutuelle

À partir des cardinalités $n_{ij} =\mid U^i \bigcap V^j \mid$ calculées pour toutes les paires de clusters des deux partitions $U$ et $V$, on peut considérer que $\frac{n_{ij}}{n}$ est une estimation de la distribution conjointe (discrète) des clusters $U^i$ et $V^j$. De même, $\frac{n_{\bullet j}}{n}$ et $\frac{n_{i\bullet}}{n}$ estiment les distributions marginales des clusters $U^i$ et $V^j$. Alors, dans le cadre de la comparaison de partitions, on peut définir l'*information mutuelle* ($IM$) de deux partitions $U$ par :

$$IM(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{r} \frac{n_{ij}}{n} log\left(n\,\frac{n_{ij}}{n_{i\bullet}\,n_{\bullet j}}\right). \tag{2.46}$$

Cette similarité mesure l'information partagée par deux partitions $U$ et $V$ en évaluant ce que la connaissance de l'une apporte sur la connaissance de l'autre. Ainsi, si $U$ et $V$ sont identiques et se déterminent ainsi bien évidemment l'une l'autre, l'information mutuelle sera maximale et égale à leur entropie $H(U) = H(V)$ de $U$ et $V$, définie comme suit :

$$H(U) = -\sum_{i}^{c} \frac{n_{i\bullet}}{n} \log \frac{n_{i\bullet}}{n} \tag{2.47}$$

Plus généralement, $IM$ est bornée ainsi :

$$IM(U,V) \leq min(H(U),H(V)) \tag{2.48}$$

et peut donc être normalisée afin d'en ramener les valeurs dans l'intervalle unité. L'égalité dans (2.48) se produit lorsque l'une des partitions comparées est un raffinement de l'autre. Dans ce cas, la normalisation par la borne décrite par (2.48) amène l'indice $IM$ à prendre sa valeur maximale de 1.

**Exemple 17.** *Pour les deux partitions $U_h$ et $V_h$ données à la Section 2.5 dont on rappelle la matrice de contingence $N(U_h,V_h) = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ qui regroupe l'ensemble des termes $n_{ij}$, l'information mutuelle est :*

$$\begin{aligned} IM(U_h,V_h) =& \frac{1}{5} log\left(5\,\frac{1}{3\times 1}\right) + \frac{2}{5} log\left(5\,\frac{2}{3\times 2}\right) + \frac{0}{5} log\left(5\,\frac{0}{3\times 2}\right) \\ &+ \frac{0}{5} log\left(5\,\frac{0}{2\times 1}\right) + \frac{0}{5} log\left(5\,\frac{0}{2\times 2}\right) + \frac{0}{5} log\left(5\,\frac{2}{2\times 2}\right) \\ =& \; 0.673 \end{aligned}$$

*En normalisant cette valeur par $min(H(U),H(V)) = H(U) = 0.673$, on obtient $IM(U_h,V-h) = 1$ puisque $V_h \subset U_h$.*

#### 2.5.2.2.b Variation d'information

Selon la même idée, Meilă [2007] propose une mesure de distance entre deux partitions fondée sur les mêmes outils de la théorie de l'information. Cette métrique, nommée *variation d'information (VI)*, se définit comme suit :

$$VI(U,V) = 2H(U,V) - H(U) - H(V) \tag{2.49}$$

où $H(U,V)$ désigne l'entropie conjointe des deux partitions, définie telle que :

$$H(U,V) = -\sum_{i}^{c}\sum_{j}^{r}\frac{n_{ij}}{n}\log\frac{n_{ij}}{n} \tag{2.50}$$

Notons que cette mesure est strictement équivalente à la métrique INFOTWO proposée par Arabie et Boorman [1973].

**Exemple 18.** *Prenons encore les deux partitions-exemple $U_h$ et $V_h$ de la Section 2.5. Selon la variation d'information, leur discordance est de :*

$$VI(U_h, V_h) = 0.382.$$

*Cette faible valeur témoigne d'une certaine proximité entre $U_h$ et $V_h$.*

## 2.6 Conclusion

Dans ce chapitre, nous avons posé les définitions et concepts clés de la comparaison de deux partitions, discuté certains enjeux et présenté un aperçu des mesures strictes historiques. Historiques car les premiers travaux dans ce domaine remontent aux années soixante-dix [Rand, 1971; Brennan et Light, 1974]. Concernant le cadre d'étude, nous nous sommes efforcés d'établir un panorama des divers domaines concernés par cette problématique afin d'en cerner au mieux les enjeux.

Pour les mesures de comparaison, toutes les mesures existantes n'ont bien évidemment pas été présentées tant la littérature sur le sujet est abondante, mais les principales approches ont été décrites et intégrées à une taxonomie différenciant les approches orientées individus et celles orientées clusters. Bien que d'autres choix étaient possibles, il est justifié par la forte implication que cette orientation peut avoir sur la notion même de la concordance entre les partitions. Nous nous sommes de plus attachés à mettre en avant ces divergences en distinguant au sein même de ces deux catégories les différentes constructions menant encore à plusieurs acceptions de ce qu'est la concordance entre deux partitions. Cette position présente le double avantage d'offrir un nouveau point de vue sur une littérature bien connue mais très

dispersée, et de poser un cadre pour l'état de l'art des mesures de comparaison de partitions non-strictes, objet de nos travaux et sujet du chapitre suivant.

# CHAPITRE 3 :

## Indices de comparaison de partitions non-strictes : état de l'art

**Résumé :** *Ce chapitre est consacré à un état de l'art des mesures de concordance et de discordance pour la comparaison de partitions non-strictes. Y sont définis les outils communs, puis différentes propositions d'extensions de mesures strictes issues de la littérature sont traitées selon la même approche taxonomique que celle adoptée au chapitre précédent. Enfin sont revues les mesures fondées sur des approches directes ne reposant sur aucune des approches historiques, qualifées de natives.*

Depuis quelques années, on a pu observer un vif regain d'intérêt pour la problématique de la comparaison de partitions, et plus particulièrement de partitions non-strictes grâce à l'essor de la logique floue et de la théorie des sous-ensembles flous. De nombreuses propositions ont vu le jour, étendant pour certaines des propositions historiques, établissant de nouvelles approches originales pour d'autres. Dans le premier cas, nous parlerons d'*extensions* de mesures strictes, tandis que dans le second on parle de mesures *natives*. Lorsqu'une mesure non-stricte, qu'elle étende une mesure historique ou qu'elle soit fondée sur une approche native, se ramène à une mesure stricte dans le cas particulier où elle est utilisée pour comparer deux partitions de $\mathbb{M}_{hcn}$, nous parlerons alors de *généralisation*. Comme nous le verrons plus tard, toutes les extensions ne sont pas nécessairement des généralisations.

Dans ce chapitre, on se propose de dresser un état de l'art espéré exhaustif de ces nouvelles mesures de comparaison non strictes. Pour illustrer par l'exemple chacune d'entre elles, nous nous réfèrerons aux trois partitions floues suivantes :

$$U_f = \begin{array}{c} \begin{matrix} \mathbf{u}_1^f & \mathbf{u}_2^f & \mathbf{u}_3^f & \mathbf{u}_4^f & \mathbf{u}_5^f \end{matrix} \\ \begin{pmatrix} 0.9 & 0.2 & 0.8 & 0.7 & 0.3 \\ 0.1 & 0.8 & 0.2 & 0.3 & 0.7 \end{pmatrix} \end{array} \begin{matrix} U_f^1 \\ U_f^2 \end{matrix}$$

$$V_f = \begin{array}{c} \begin{matrix} \mathbf{v}_1^f & \mathbf{v}_2^f & \mathbf{v}_3^f & \mathbf{v}_4^f & \mathbf{v}_5^f \end{matrix} \\ \begin{pmatrix} 0.8 & 0.1 & 0.2 & 0.3 & 0.2 \\ 0.1 & 0.2 & 0.7 & 0.6 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 & 0.7 \end{pmatrix} \end{array} \begin{matrix} V_f^1 \\ V_h^2 \\ V_f^3 \end{matrix}$$

$$W_f = \begin{array}{c} \begin{matrix} \mathbf{w}_1^f & \mathbf{w}_2^f & \mathbf{w}_3^f & \mathbf{w}_4^f & \mathbf{w}_5^f \end{matrix} \\ \begin{pmatrix} 0.7 & 0.1 & 0.2 & 0.2 & 0 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.3 \\ 0.1 & 0.1 & 0.7 & 0.6 & 0.1 \\ 0.1 & 0.2 & 0 & 0.1 & 0.6 \end{pmatrix} \end{array} \begin{matrix} W_f^1 \\ W_f^2 \\ W_f^3 \\ W_f^4 \end{matrix}$$

conçues de telle sorte que les partitions max-défuzzifiées correspondantes soient respectivement les partitions strictes $U_h$, $V_h$ et $W_h$ de la page 18. À l'instar de leur pendant strict, elles sont de plus des raffinements stricts les unes des autres ($W_f \subset V_f \subset U_f$) au sens de la Définition 5 (page 9).

## 3.1 Outils préliminaires

### 3.1.1 Normes et conormes triangulaires

Menger [1942] est le premier à définir une *norme triangulaire* (ou *t-norme*) dans le but de d'étendre l'inégalité triangulaire des espaces métriques classiques aux espaces métriques probabilistes. Dans le contexte qui nous intéresse, et c'est l'acception actuelle que nous retiendrons, une t-norme permet d'étendre à l'intervalle unité $[0,1]$, à la fois l'opérateur booléen de conjonction $\wedge$ et l'intersection ensembliste classique, comme proposé par Schweizer et Sklar [1983].

**Définition 15.** *[Schweizer et Sklar, 1983] Une t-norme est un opérateur* $\top : [0,1]^2 \to [0,1]$, *tel que, pour tout* $a,\ b,\ c \in [0,1]$ *:*

(II-1) $\top(a,b) = \top(b,a)$ (commutativité),

(II-2) $\top(a,\top(b,c)) = \top(\top(a,b),c)$ (associativité),

(II-3) $\top(a,b) \leq \top(a,c)$ *si* $b \leq c$ (monotonie),

(II-4) $\top(a,1) = 1$ (élément neutre).

Les opérateurs minimum et produit, satisfaisant l'ensemble de ces axiomes, sont ainsi les deux plus connues des t-normes dites basiques données par la Table 3.1.

TABLE 3.1 – Normes et conormes triangulaires basiques

| $\top$-norme | $\top(a,b)$ | $\perp(a,b)$ |
|---|---|---|
| Standard | $\top_M(a,b) = \min(a,b)$ | $\perp_M(a,b) = \max(a,b)$ |
| Produit | $\top_P(a,b) = a\,b$ | $\perp_P(a,b) = a+b-a\,b$ |
| Łukasiewicz | $\top_Ł(a,b) = \max(a+b-1,0)$ | $\perp_Ł(a,b) = \min(a+b,1)$ |
| Drastique | $\top_D(a,b) = \begin{cases} 0 \text{ si } (a,b) \in [0,1[^2 \\ \min(a,b) \text{ sinon} \end{cases}$ | $\perp_D(a,b) = \begin{cases} 1 \text{ si } (a,b) \in ]0,1]^2 \\ \max(a,b) \text{ sinon} \end{cases}$ |

Par ailleurs, la t-norme minimum ou t-norme *standard* $\top_M$ borne supérieurement l'ensemble des normes triangulaires, tel que :

$$\top(a,b) \leq min(a,b) = \top_M(a,b). \tag{3.1}$$

La Figure 3.1 présente les iso-surfaces obtenues pour les quatre t-normes basiques, où plus la couleur est chaude, plus la valeur de sortie de l'opérateur est élevée.

À toute t-norme on associe un opérateur dual étendant à la fois l'opérateur booléen de disjonction $\vee$ et l'union ensembliste classique. Cet opérateur est appelé *conorme triangulaire* ou *t-conorme*.

**Définition 16.** *Une t-conorme est un opérateur* $\perp : [0,1]^2 \to [0,1]$ *commutatif, associatif et monotone qui accepte* 0 *comme élément neutre.*

On peut exprimer une t-conorme à partir de sa t-norme duale :

$$\perp(a,b) = N(\top(N(a),N(b))), \tag{3.2}$$

où $N(a)$ est le complément de $a$ au sens d'une *négation floue*, si bien que tout couple $(\top,\perp)$ de t-norme et t-conorme duales généralise au cas flou les lois de De Morgan.

**Définition 17.** *Une fonction* $N : [0,1] \to [0,1]$ *est une* négation floue *si elle satisfait :*

*(*III-1*)* $N(0) = 1$ *et* $N(1) = 0$ *(*conditions aux bornes*),*

*(*III-2*)* $a \leq b \Rightarrow N(a) \geq N(b)$ *(*monotonie*).*

*Une négation floue est dite* forte *si elle est involutive, c'est-à-dire qu'elle satisfait de plus :*

*(*III-3*)* $N(N(a)) = a$ *(involution).*

La stricte négation floue $\overline{a} = 1 - a$ est ainsi une négation floue forte, et c'est celle qui sera considérée pour la suite de ce mémoire.

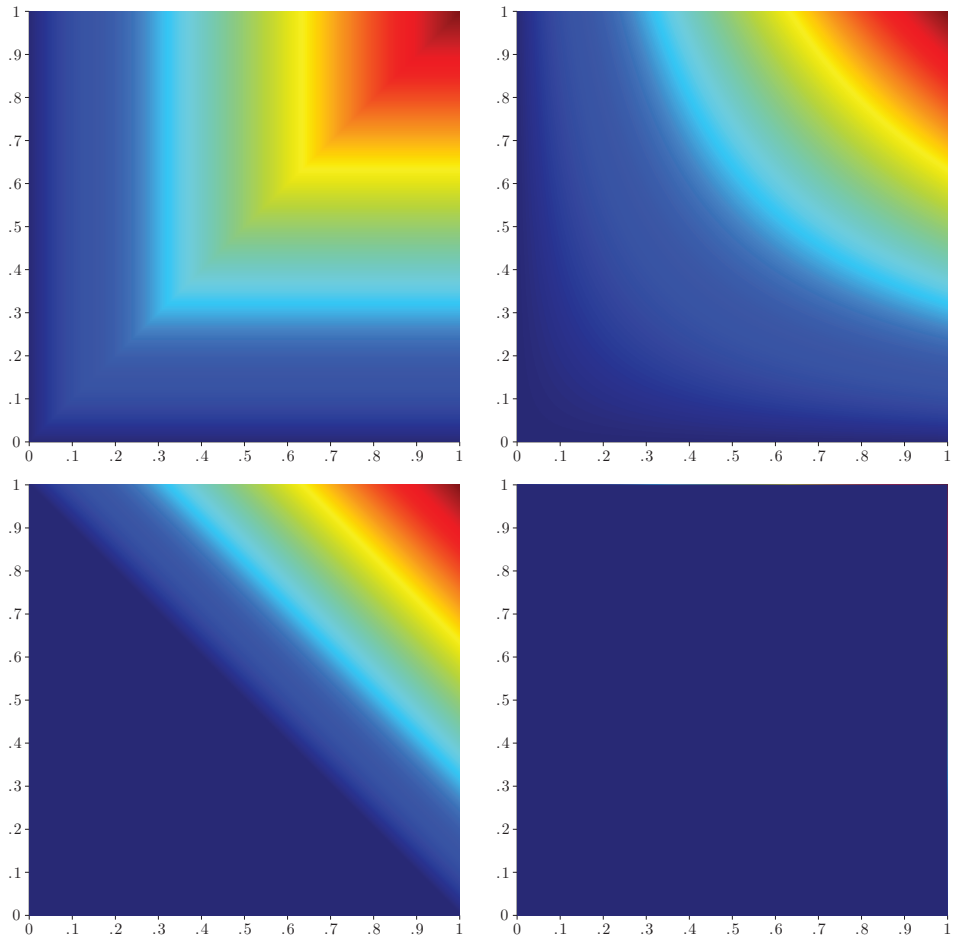


FIGURE 3.1 – Première ligne : iso-surfaces des t-normes basiques $\top_M$, $\top_P$. Deuxième ligne : iso-surfaces des t-normes basiques $\top_Ł$ et $\top_D$.

Les conormes associées à chacune des t-normes basiques sont encore données par la Table 3.1. Remarquons que la t-conorme duale de $\top_M$ est l'opérateur maximum qui borne inférieurement l'ensemble des t-conormes, si bien que l'on a, $\forall a, b \in [0,1]$ :

$$\top(a,b) \leq \min(a,b) \leq \max(a,b) \leq \bot(a,b). \tag{3.3}$$

Les iso-surfaces des conormes triangulaires basiques sont représentées à la Figure 3.2. Plus la couleur est chaude, plus la valeur de sortie de l'opérateur est élevée.

Il existe un grand nombre de couples de t-normes et t-conormes duales associées [Klement et Mesiar, 2005], et parmi les plus populaires on distinguera des familles de couples dites paramétriques, dont les pricipales normes triangulaires sont données par la Table 3.2 ci-après[1]. La particularité de ces familles est de permettre

1. Seules les normes triangulaires sont données par souci de concision, leur t-conorme duale respective pouvant aisément être retrouvée par 3.2 en utilisant la négation stricte.

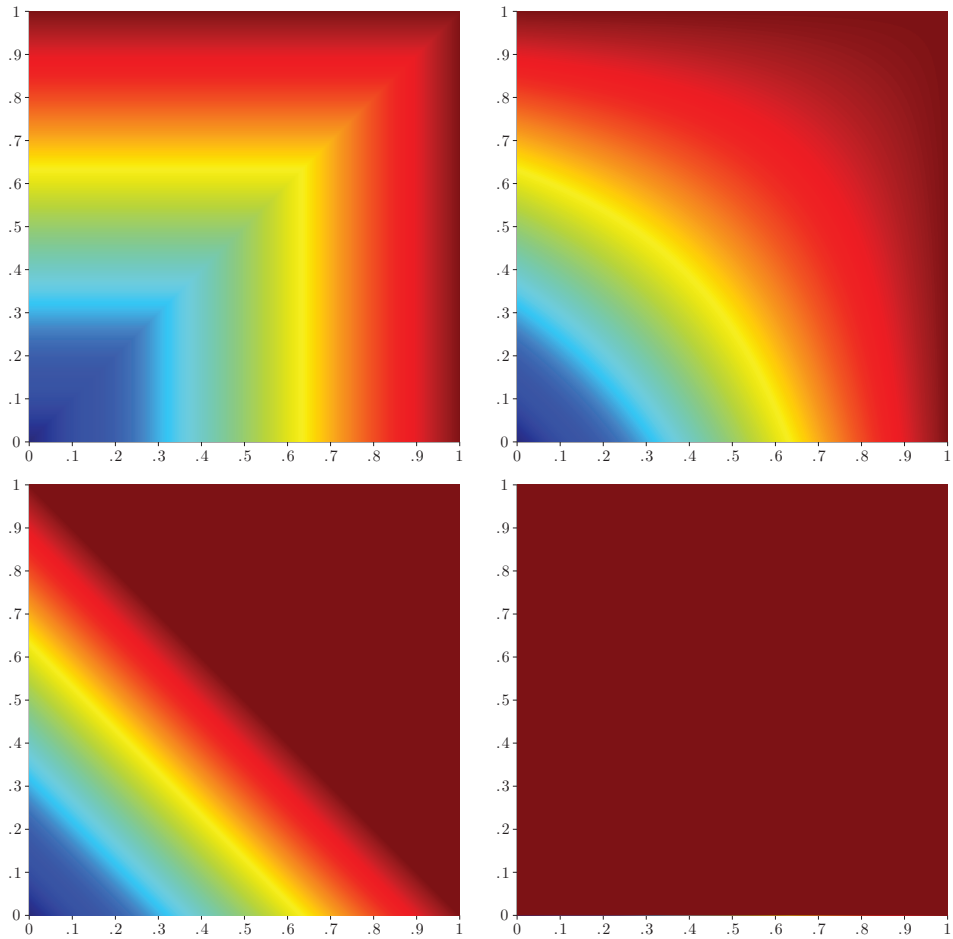


FIGURE 3.2 – Première ligne : iso-surfaces des t-conormes basiques $\top_M$, $\top_P$. Deuxième ligne : iso-surfaces des t-conormes basiques $\top_Ł$ et $\top_D$.

au praticien de contrôler, en fonction d'un paramètre $\lambda$, la manière dont les valeurs $a$ et $b$ sont agrégées. Comme l'illustre la Figure 3.3 pour la famille de Hamacher, de telles t-normes et t-conormes peuvent ainsi être adaptées de sorte d'adopter un comportement plus ou moins drastique, c'est-à-dire tendre ou non vers les normes $\top_D$ et $\bot_D$. Elle présentent aussi l'intérêt de se ramener à des couples $(\top, \bot)$ basiques pour certaines valeurs particulières de $\lambda$. Par exemple, les couples $(\top_{AA_\lambda}, \bot_{AA_\lambda})$ et $(\top_{H_\lambda}, \bot_{H_\lambda})$ d'Aczel-Alsina et de Hamacher se réduisent au couple produit $(\top_P, \bot_P)$ en choisissant $\lambda = 1$. De la même manière, les couples de Schweizer-Sklar et de Yager sont égaux à celui de Łukasiewicz pour $\lambda = 1$.

TABLE 3.2 – Principales familles paramétriques de normes triangulaires

| $\top$-norme | $\top(a,b)$ | | |
|---|---|---|---|
| Aczel-Alsina | $\top_{AA_\lambda} =$ | $\top_D(a,b)$ | si $\lambda = 0$ |
| | | $\top_M(a,b)$ | si $\lambda = +\infty$ |
| | | $e^{-\left((-\ln a)^\lambda + (-\ln b)^\lambda\right)^{1/\lambda}}$ | si $\lambda \in ]0,+\infty[$ |
| Dombi | $\top_{D_\lambda} =$ | $\top_D(a,b)$ | si $\lambda = 0$ |
| | | $\top_M(a,b)$ | si $\lambda = +\infty$ |
| | | $\left(1 + \left(\left(\frac{1-a}{a}\right)^\lambda + \left(\frac{1-b}{b}\right)^\lambda\right)^{1/\lambda}\right)^{-1}$ | si $\lambda \in ]0,+\infty[$ |
| Frank | $\top_{F_\lambda} =$ | $\top_M(a,b)$ | si $\lambda = 0$ |
| | | $\top_P(a,b)$ | si $\lambda = 1$ |
| | | $\top_Ł(a,b)$ | si $\lambda = +\infty$ |
| | | $log_\lambda\left(1 + \frac{(\lambda^a-1)(\lambda^b-1)}{\lambda-1}\right)$ | si $\lambda \in ]0,1[\cup]1,+\infty[$ |
| Hamacher | $\top_{H_\lambda} =$ | $0$ | si $\lambda = a = b = 0$ |
| | | $\top_D(a,b)$ | si $\lambda = +\infty$ |
| | | $\frac{a\,b}{(\lambda+(1-\lambda)(a+b-a\,b))}$ | si $\lambda \in [0,+\infty[$ |
| Schweizer-Sklar | $\top_{SS_\lambda} =$ | $\top_M(a,b)$ | si $\lambda = -\infty$ |
| | | $\top_P(a,b)$ | si $\lambda = 0$ |
| | | $\top_D(a,b)$ | si $\lambda = +\infty$ |
| | | $\left(\max\left(a^\lambda + b^\lambda - 1, 0\right)\right)^{1/\lambda}$ | si $\lambda \in ]-\infty,0[\cup]0,+\infty[$ |
| Sugeno-Weber | $\top_{SW_\lambda} =$ | $\top_D$ | si $\lambda = -1$ |
| | | $\top_P(a,b)$ | si $\lambda = +\infty$ |
| | | $\max\left(\frac{a+b-1+\lambda ab}{1+\lambda}, 0\right)$ | si $\lambda \in ]-1,+\infty[$ |
| Yager | $\top_{Y_\lambda} =$ | $\top_D$ | si $\lambda = 0$ |
| | | $\top_M(a,b)$ | si $\lambda = +\infty$ |
| | | $\max\left(1 - \left((1-a)^\lambda + (1-b)^\lambda\right)^{1/\lambda}, 0\right)$ | si $\lambda \in ]0,+\infty[$ |

**Définition 18.** *Une t-norme continue est dite* archimédienne *si elle satisfait la propriété suivante :*

$$\top(a,a) < a. \tag{3.4}$$

*De la même manière, une t-conorme continue est archimédienne si*

$$\perp(a,a) > a. \tag{3.5}$$

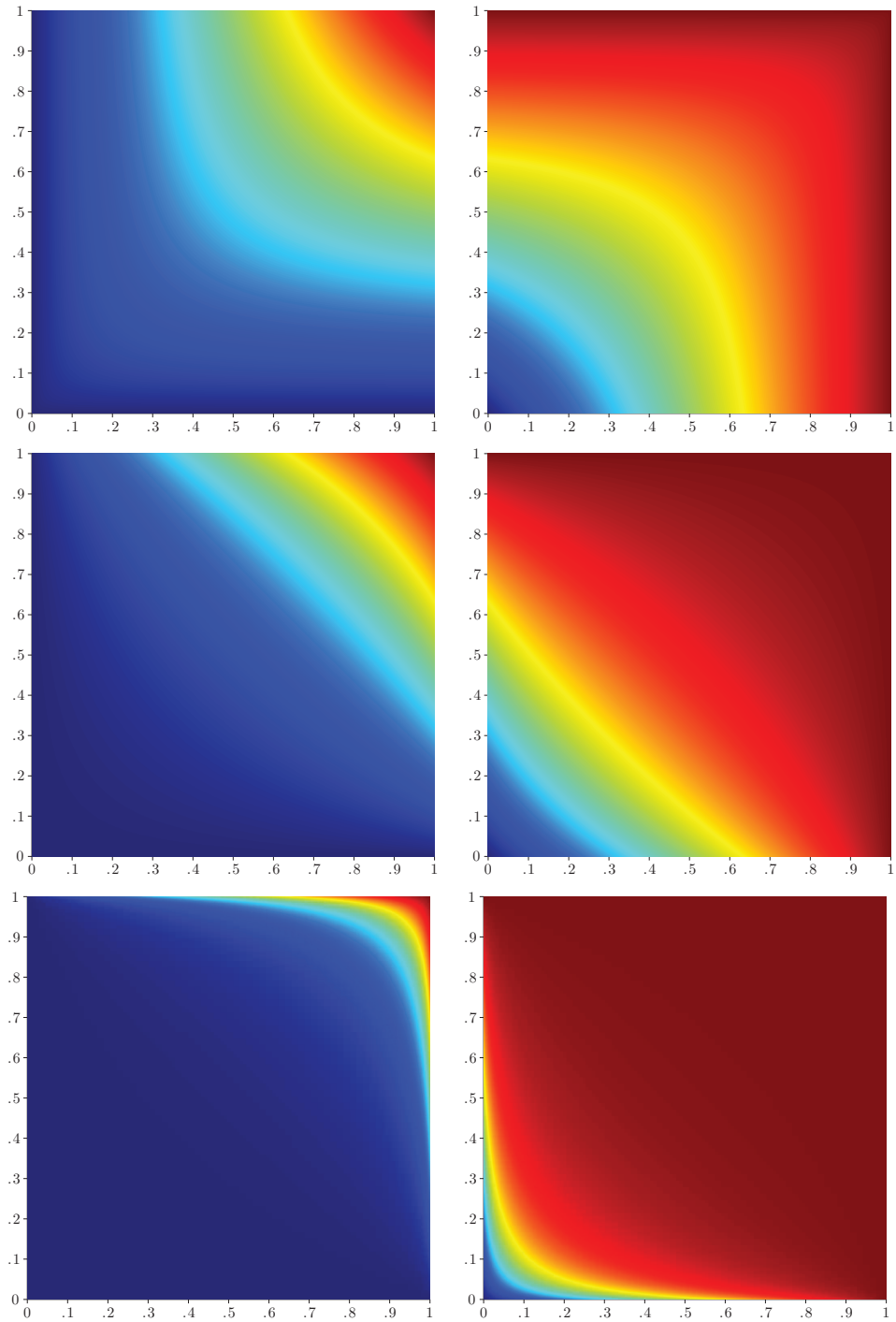


FIGURE 3.3 – Première colonne : iso-surfaces de $\top_{H_\lambda}$ pour $\lambda = 0$, $\lambda = 5$ et $\lambda = 100$. Deuxième colonne : $\perp_{H_\lambda}$ pour $\lambda = 0$, $\lambda = 5$ et $\lambda = 100$.

**Définition 19.** *Pour toute t-norme $\top$ continue archimédienne, il existe une fonction $f_\top : [0,1] \to [0,\infty[$, continue décroissante, et telle que $f_\top(1) = 0$, appelée* générateur additif *de $\top$, permettant de représenter cette dernière comme suit :*

$$\top(a,b) = f_\top^{(-1)}\big(f_\top(a) + f_\top(b)\big), \tag{3.6}$$

*et plus généralement (cas n-aire) :*

$$\top_{i=1}^{n} a_i = f_\top^{(-1)}\Big(\sum_{i=1}^{n} f_\top(a_i)\Big), \tag{3.7}$$

*avec $f_\top^{(-1)}$, la pseudo-inverse de $f$ définie tel que :*

$$f_\top^{(-1)}(a) = \begin{cases} f_\top^{-1}(a) & si\ 0 \leq a < f_\top(0) \\ 0 & sinon \end{cases} \tag{3.8}$$

*où $f^{-1}$ est la fonction inverse de $f$.*

*De même, toute t-norme $\top$ continue archimédienne possède un* générateur multiplicatif *$g_\top : [0,1] \to [0,1]$, strictement croissant, et tel que $g_\top(1) = 1$, de sorte que $\top$ puisse être exprimé par :*

$$\top(a,b) = g_\top^{(-1)}\big(g_\top(a)\ g_\top(b)\big). \tag{3.9}$$

*et, dans le cas n-aire :*

$$\top_{i=1}^{n} a_i = g_\top^{(-1)}\Big(\prod_{i=1}^{n} g_\top(a_i)\Big). \tag{3.10}$$

*où $g_\top^{(-1)}$ est la pseudo-inverse de $g$.*

À l'exception de la t-norme minimum $\top_M$, toutes les t-normes présentées dans les Tables 3.1 et 3.2 sont archimédiennes et il en va de même pour chacune de leurs conormes associées. Leurs générateurs additifs ainsi que leur inverse $f^{-1}$ sont donnés par la Table 3.3. Précisons que dans le cas des t-normes paramétriques pour lesquelles certaines valeurs particulières de leur paramètre $\lambda$ les ramènent à l'une des t-normes basiques, c'est le générateur de cette dernière qui sera utilisé.
L'utilisation de générateurs a un double intérêt. Le premier, dont nous n'avons pas besoin pour nos travaux, est calculatoire. En effet, comme illustré à la Figure3.4 pour la t-norme $\top_{H_5}$, le temps de calcul pour un nombre important d'opérandes est largement diminué. Le second, sur lequel nous appuierons une proposition au Chapitre 4, page 85, est formel.

TABLE 3.3 – Générateurs additifs et leur inverse pour quelques t-normes archimédiennes

| ⊤-norme | générateur additif $f(a)$ | inverse $f^{-1}(a)$ |
|---|---|---|
| Produit | $-\ln a$ | $e^{-a}$ |
| Łukasiewicz | $1-a$ | $-(a-1)$ |
| Aczel-Alsina | $(-\ln a)^{\lambda}$ | $e^{-a^{1/\lambda}}$ |
| Dombi | $\left(\frac{1-a}{a}\right)^{\lambda}$ | $\left(1+a^{1/\lambda}\right)^{-1}$ |
| Frank | $\ln\left(\frac{\lambda-1}{\lambda^a-1}\right)$ | $\frac{\ln(1+(\lambda-1)\,e^{-a})}{\ln\lambda}$ |
| Hamacher | $\ln\left(\frac{\lambda+(1-\lambda)a}{a}\right)$ | $\frac{\lambda}{e^a+\lambda-1}$ |
| Schweizer-Sklar | $\frac{1-a^{\lambda}}{\lambda}$ | $(1-\lambda a)^{1/\lambda}$ |
| Sugeno-Weber | $1-\frac{\ln(1+\lambda a)}{\ln(1+\lambda)}$ | $\frac{(1+\lambda)^{1-a}-1}{\lambda}$ |
| Yager | $(1-a)^{\lambda}$ | $1-a^{1/\lambda}$ |

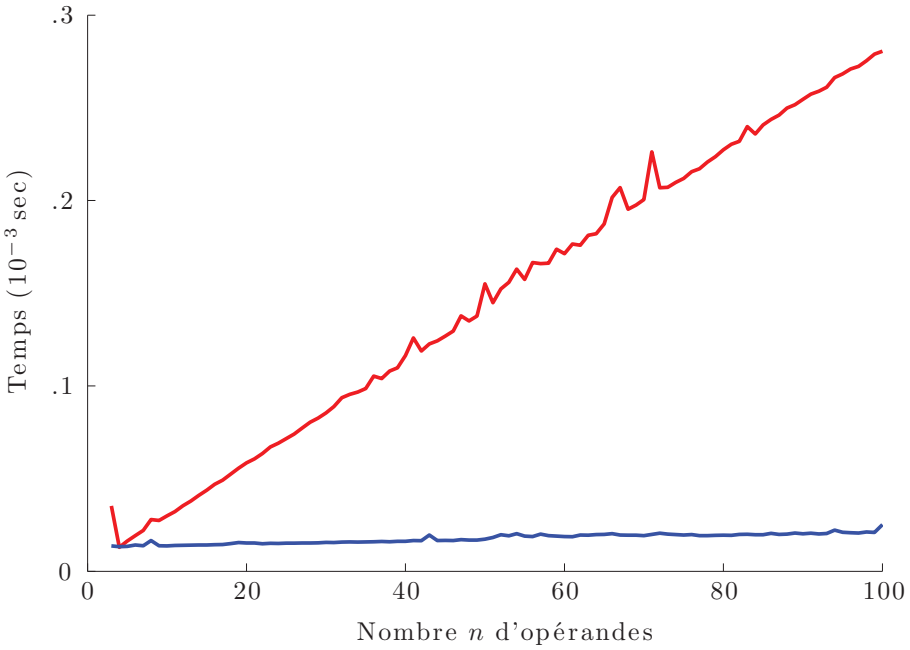



FIGURE 3.4 – Temps de calcul de la t-norme $\top_{H_5}$ avec (en bleu) et sans (en rouge) utilisation de son générateur additif, en fonction du nombre d'opérandes

### 3.1.2 Implications floues

À partir des connecteurs logiques flous que sont les normes et conormes triangulaires décrits plus tôt à la page 42, la logique floue s'est dotée d'outils pour la construction de nouveaux systèmes d'inférence. Parmi ceux-ci, l'implication floue, qui généralise l'implication stricte bien connue, est largement utilisée. Elle permet de modéliser des propositions conditionnelles de type *si $a$ alors $b$*, où $a$ et $b$ sont des prédicats flous.

**Définition 20.** *Une implication floue est une fonction $\mathcal{I} : [0,1]^2 \to [0,1]$ , $(a,b) \mapsto \mathcal{I}(a,b)$ telle que :*

(IF-1) *$\mathcal{I}(a,b)$ est non-croissante avec la première variable $a$ et non-décroissante avec la seconde variable $b$,*

(IF-2) *$\mathcal{I}(0,0) = \mathcal{I}(1,1) = 1$,*

(IF-3) *$\mathcal{I}(1,0) = 0$,*

*de sorte que sur $\{0,1\}^2$, elle coïncide avec l'implication stricte.*

Une implication floue peut par ailleurs satisfaire de nombreuses propriétés additionnelles, parmis lesquelles on recense [Baczyński et Jayaram, 2008] :

(*IF-4*) $\mathcal{I}(a,a) = 1,\ \forall a \in [0,1]$ (*principe d'identité*),

(*IF-5*) $\mathcal{I}(1,a) = a,\ \forall a \in [0,1]$ (*principe de bord*),

(*IF-6*) $\mathcal{I}(a,\mathcal{I}(b,c)) = \mathcal{I}(b,\mathcal{I}(x,c))$ (*principe d'échange*),

(*IF-7*) $\mathcal{I}(a,b) = 1 \Leftrightarrow a \leq b$ (*principe de confinement*),

(*IF-8*) $\mathcal{I}(a,b) = \mathcal{I}(N(a),N(b))$, où $N$ est une négation floue (*principe de contraposition*).

On relève cinq types d'implications floues, que sont [Mas et al., 2007] :

- les *S-implications*, définies par :

$$\mathcal{I}_{\perp}(a,b) = \perp(1-a,b), \tag{3.11}$$

  qui généralisent directement l'implication stricte, définie telle que $a \to b = \overline{a} \vee b$.
- les *Quantum mechanic Logic implications*, ou *QL-implications*, définies par :

$$\mathcal{I}_{QL}(a,b) = \perp(\overline{a}, \top(a,b)), \tag{3.12}$$

- les *D-implications*, définies par :

$$\mathcal{I}_{D}(a,b) = \perp(\top(\overline{a},\overline{b}),b), \tag{3.13}$$

  et qui sont la contraposée des QL-implications, autrement dit $\mathcal{I}_D(a,b) = \mathcal{I}_{QL}(\overline{b},\overline{a})$.

– les *A-implications*, définies par :

$$\mathcal{I}_A(a,b) = b^a, \tag{3.14}$$

– les *implications résiduelles*, ou *R-implications*, définies par :

$$\mathcal{I}_\top(a,b) = \sup_t\{t \in [0,1] \ : \ \top(a,t) \leq b\}. \tag{3.15}$$

Toutes les R-implications satisfont par construction les propriétés *(IF-4)*, *(IF-5)* et *(IF-6)*. Par ailleurs, si la t-norme utilisée pour leur construction est continue à gauche, alors *(IF-7)* est aussi satisfaite. Ainsi, à l'exception de celle construite à partir de la t-norme drastique $\top_D$ qui n'est pas continue à gauche, toutes les implications résiduelles dérivées des t-normes présentées par les Tables 3.1 et 3.2 satisfont les propriétés *(IF-4)*-*(IF-7)*. La propriété *(IF-8)* n'est généralement pas satisfaite pour la négation stricte $N(a) = \overline{a}$. Les principales implications résiduelles, dérivées des normes triangulaires basiques et de quelques familles de t-normes paramétriques [Fono et al., 2007; Le Capitaine et Frélicot, 2009] sont recensées dans la Table 3.4. Au Chapitre 4, page 112, nous proposerons de fonder la construction d'un indice original de comparaison de partitions floues selon une approche contingentielle, entre autres sur des R-implications.

Table 3.4 – Principales implications résiduelles

| $\top$-norme | $\mathcal{I}(a,b)$ |
|---|---|
| Minimum (Gödel) | $\mathcal{I}_M(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ b & \text{sinon} \end{cases}$ |
| Produit (Goguen) | $\mathcal{I}_P(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ \frac{b}{a} & \text{sinon} \end{cases}$ |
| Łukasiewicz | $\mathcal{I}_Ł(a,b) = min(1, 1-a+b)$ |
| Hamacher | $\mathcal{I}_{H_\lambda}(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ \frac{b(\lambda+a-\lambda a)}{b(\lambda+a-\lambda a)+a-b} & \text{sinon} \end{cases}$ |
| Dombi | $\mathcal{I}_{D_\lambda}(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ \left(1+\left(\left(\frac{1-b}{b}\right)^\lambda - \left(\frac{1-a}{a}\right)^\lambda\right)^{\frac{1}{\lambda}}\right)^{-1} & \text{sinon} \end{cases}$ |
| Yager | $\mathcal{I}_{Y_\lambda}(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ 1-((1-b)^\lambda - (1-a)^\lambda)^{\frac{1}{\lambda}} & \text{sinon} \end{cases}$ |
| Frank | $\mathcal{I}_{F_\lambda}(a,b) = \begin{cases} 1 & \text{si } b \geq a \\ \log_\lambda(1+\frac{(\lambda^b-1)(\lambda-1)}{\lambda^a-1}) & \text{sinon} \end{cases}$ |

### 3.1.3 Mesures de compatibilité entre ensembles flous

En s'appuyant sur les travaux de Tversky [1977], Bouchon-Meunier et al. [1996] proposent d'unifier les extensions aux ensembles flous des mesures de compatibilité entre ensembles stricts (Définition 13, page 32).
Dans la suite, on notera $\mathcal{F}(X)$ l'ensemble de tous les ensembles flous définis sur $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$.

**Définition 21.** *Une* mesure de compatibilité *entre deux sous-ensembles flous $A$ et $B$ de $\mathcal{F}(X)$ est une fonction $M(A,B) = F(a,b,c)$, avec $a = f(\top(A,B))$, $b = f(A \ominus B)$ et $c = f(B \ominus A)$, où $\top$ est une t-norme et $f$ est une mesure floue*[2] [3]*, et $F$ est une fonction :* $\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow [0,1]$.

Selon les propriétés de $F$, on distingue plusieurs types de mesures de compatibilité :

- une *mesure de similitude* si $F$ est non décroissante avec $a$ et non croissante ni avec $b$, ni avec $c$,
- une *mesure de satisfiabilité* si $F$ est croissante avec $a$, décroissante avec $b$, indépendante de $c$ et telle que $F(a,0,c) = 1$ et $F(0,b,c) = 0$,
- une *mesure de ressemblance* si $F$ est croissante avec $a$, décroissante avec $b$ et $c$ et telle que $F(a,0,0) = 1$ et $F(a,b,c) = F(a,c,b)$,
- une *mesure de dissimilarité* si $F$ est indépendante de $a$, croissante avec $b$ et $c$ et si $M$ est telle que $M(A,B) = 0$.

Pour plus de détails, ainsi que pour une étude du comportement de chacune de ces familles de mesures de compatibilité à partir de leur pouvoir discriminant, le lecteur pourra se référer à [Bouchon-Meunier et al., 1996] et [Rifqi et al., 2000].

## 3.2 Différentes extensions de mesures strictes

On entend par *extension* d'une mesure stricte toute mesure fondée sur la redéfinition des équations d'une mesure vouée originellement à la comparaison de partitions strictes, telle que celles décrites au Chapitre 2, afin d'en élargir la portée aux cas des partitions non-strictes de $\mathbb{M}_{scn}$. Ces extensions peuvent être conçues de telle sorte qu'elles se ramènent directement à leur pendant strict lorsque les partitions à comparer appartiennent à $\mathbb{M}_{hcn}$ (indices de Campello [2007], page 53 ; indices de Anderson

2. Une mesure $f$ est une mesure floue si :
- $f(A) \leq f(B)$ lorsque $A \subseteq B$
- $f(\emptyset) = 0$ et $f(X) = 1$

3. Dans le cadre qui nous intéresse, $f$ est une mesure cardinale. On prendra donc la *cardinalité floue* (ou $\Sigma$-count) définie par : $f_\Sigma(A) = \sum_{x \in X} a(x)$, où $a$ est la fonction d'appartenance à $A$

et al. [2010], page 56 ; indices de Borgelt [2006], page 60 ; indices de Brouwer [2009a], page 63 ; distance de transfert floue, page 64), mais ce n'est pas toujours le cas (indices de Ceccarelli et Maratea [2008], page 58). Nous nous attachons à revoir les différentes extensions de mesures strictes de la littérature, selon la même approche taxonomique que celle adoptée au Chapitre 2. Sont donc d'abord passées en revues les extensions de mesures orientées individus, puis les extensions de mesures orientées clusters.

### 3.2.1 Approches orientées individus

Les mesures décrites dans cette section proposent d'étendre aux cas non-stricts tout ou partie des mesures d'appariement revues à la page 18. Ainsi, ces méthodes sont fondées sur une redéfinition de la matrice des contingences-paire en s'appuyant sur les différentes approches calculatoires (ensembliste, contingentielle, coincidentielle) présentées au Chapitre 2 aux pages (25, 28, 29). À partir de ces matrices de contigence-paires non-strictes toutes les mesures d'appariements peuvent être étendues aux partitions de $\mathbb{M}_{scn}$, parmi lesquelles celles recensées dans la Table 2.3, page 26.

Toutefois, nous pensons que les mesures fondées sur des hypothèses statistiques tels que les indices corrigés pour la chance ne devraient être considérés dans ce cas qu'avec précaution, voire tout simplement oubliés pour la comparaison de partitions non-strictes puisque les hypothèses sur lesquelles elles reposent ne tiennent généralement plus dans le cas des partitions floues ou possibilistes. Citons pour l'exemple le Rand Ajusté (2.15) discuté page 24 qui ne peut clairement être étendu aux cas non-stricts sans mettre à mal l'hypothèse qui justifie sa construction, à savoir que les termes de la matrice de contingence croisant deux partitions aléatoires suivent une loi hypergéométrique – discrète par définition.

#### 3.2.1.1 Approche ensembliste : indices de Campello

Campello [2007] propose une première approche pour la définition d'une table des contingence-paires de deux partitions non-strictes $U$ et $V$, étendant ainsi aux cas flou et possibiliste l'intégralité des mesures d'appariement revues au Chapitre 2. Cette approche consiste en une reformulation des équations (2.16), (2.17) et (2.18) en remplaçant chaque opérateur ensembliste par sa généralisation floue. Ainsi, les quatres termes $\{m^{\top}_{\alpha\beta} : \alpha, \beta \in \{0,1\}\}$ de la table des contingences-paire sont alors simplement donnés, pour n'importe quel couple de normes triangulaires $(\top, \bot)$ par :

$$m^{\top}_{\alpha\beta}(U,V) \quad = \quad \sum_{k=2}^{n}\sum_{l=1}^{k-1} \top(U_{\alpha}(k,l),\ V_{\beta}(k,l)) \tag{3.16}$$

avec :

$$U_1(k,l) = \bigperp_{i=1}^{c} U^{ii}(k,l) \quad \text{et} \quad V_1(k,l) = \bigperp_{i=1}^{r} V^{ii}(k,l), \tag{3.17}$$

$$U_0(k,l) = \bigperp_{\substack{i \neq j \\ i,j=1}}^{c} U^{ij}(k,l) \quad \text{et} \quad V_0(k,l) = \bigperp_{\substack{i \neq j \\ i,j=1}}^{r} V^{ij}(k,l), \tag{3.18}$$

et où $U^{ij}$ et $V^{ij}$ sont les ensembles flous des paires d'individus $\{x_k, x_l\}$ appartenant au $i$-ième cluster et au $j$-ième cluster de $U$ et de $V$ respectivement . Le degré d'appartenance de chaque paire $\{\mathbf{x}_k, \mathbf{x}_l\}$ de $X$ à ces ensembles $U^{ij}$ et $V^{ij}$ correspond à la valeur de vérité floue de la proposition "$\mathbf{x}_k$ appartient à $U^i$ (respectivement $V^i$) ET $\mathbf{x}_l$ appartient à $U^j$ (respectivement $V^j$)", et est ainsi défini tel que :

$$U^{ij}(k,l) = \top(u_{ik}, u_{jl}) \text{ et } V^{ij}(k,l) = \top(v_{ik}, v_{jl}) \tag{3.19}$$

On notera $M_C^{\top}(U,V)$ la table des contingences-paire résultante, de laquelle peuvent être dérivées et généralisées toutes les mesures d'appariement décrites au Chapitre 2 et en particulier l'indice de Rand $RI_C^{\top}$ et l'indice de Jaccard $JI_C^{\top}$.

Comme le précise Campello lui-même, et comme l'illustre l'Exemple 19, les mesures dérivées de cette proposition perdent systématiquement leur propriété de réflexivité telle que définie par *(II-1)* dans le cas de la comparaison de deux partitions non-strictes de $\mathbb{M}_{scn}$. L'auteur préconise alors de restreindre l'usage de sa proposition à la comparaison d'une partition floue ou possibiliste avec une partition stricte, limitant donc les mesures dérivées à une utilisation externe où la partition de référence est stricte. Cette perte induit par ailleurs qu'aucune des mesures d'appariement dérivées (ou son dual, le cas échéant) n'est une métrique, ni même une similarité dans l'espace des partitions de $\mathbb{M}_{scn}$.

**Exemple 19.** *Considérons les deux partitions floues $U_f$ et $V_f$ (page 42). Selon l'approche proposée par Campello, leur matrice de contingence-paires, calculée avec le couple $(\top_P, \perp_P)$, est :*

$$M_C^{\top_P}(U_f, V_f) = \begin{pmatrix} 1.294 & 2.535 \\ 1.206 & 2.995 \end{pmatrix}.$$

*À partir de celle-ci, les indices de Rand et de Jaccard ont pour valeur :*

$$RI_C^{\top_P}(U_f, V_f) = 0.534 \text{ et } JI_C^{\top_P}(U_f, V_f) = 0.257.$$

*On observe que ces valeurs sont très inférieures à celles obtenues avec les mêmes indices pour les partitions max-déffuzifiées respectives $U_h$ et $V_h$, pour lesquelles nous avions :*

$$RI(U_h, V_h) = 0.800 \text{ et } JI(U_h, V_h) = 0.500.$$

*Cette différence est due au caractère flou de $U_f$ et $V_f$, qui sont par définition moins précises que $U_h$ et $V_h$, si bien qu'elles concordent moins. Notons que cette remarque peut-être faite pour toutes les propositions détaillées dans ce chapitre.*
*La Table 3.5 donne l'ensemble des résultats obtenus avec $RI^C_{\top_P}$ et $JI^C_{\top_P}$ en croisant toutes les partitions exemple $U_f$, $V_f$ et $W_f$. D'une façon générale, on remarque que la dynamique (propriété (II-4)) est très faible pour cet exemple. Aussi, on constate que $RI_C^{\top_P}(U_f, U_f) < 1$ et que $JI_C^{\top_P}(W_f, W_f) \leq JI_C^{\top_P}(W_f, U_f)$. Ce dernier résultat contre-intuitif est directement induit par la perte de la propriété de réflexivité (II-1) de la mesure, et s'explique par le fait que les termes $m_{01}$ et $m_{10}$ présentent des valeurs plus importantes lorsque l'on compare $W_f$ avec elle-même que lorsqu'on la compare avec $U_f$, à cause du plus grand nombre de clusters dans $W_f$.*

| $RI_C^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.559 | 0.534 | 0.535 |
| $V_f$ | 0.534 | 0.600 | 0.617 |
| $W_f$ | 0.535 | 0.617 | 0.652 |

| $JI_C^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.364 | 0.257 | 0.217 |
| $V_f$ | 0.257 | 0.225 | 0.192 |
| $W_f$ | 0.217 | 0.192 | 0.178 |

TABLE 3.5 – $RI_C^{\top_P}$ (à gauche) et $JI_C^{\top_P}$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

Hüllermeier et Rifqi [2009] formulent une vive critique à l'égard de cette proposition, énonçant que l'usage des normes triangulaires amène les mesure dérivées à méconsidérer les relations topologiques existant entre les données décrites par les partitions. En donnant une acception probabiliste aux degrés d'appartenance d'une partition $U \in \mathbb{M}_{fcn}$, et en considérant le couple $(\top_P, \perp_Ł)$, ils fondent leur critique sur l'interprétation que l'on peut donner aux termes $U_1(k, l)$, selon laquelle ce terme donne la probabilité que deux individus $\mathbf{x}_k$ et $\mathbf{x}_l$ appartiennent au même cluster de $U$ sous l'hypothèse que le cluster d'appartenance de chaque individu soit choisi indépendamment de celui de l'autre. Une telle hypothèse entre alors en conflit avec les relations spatiales que l'on peut constater dans l'espace de représentation de $X$, puisque l'on s'attend à ce que deux individus très proches, presque confondus, appartiennent systématiquement au même cluster si bien que l'appartenance de l'un à un cluster conditionne celle de l'autre. En s'appuyant sur ce point, ils poursuivent leur critique en défendant le point de vue selon lequel $U_1(k, l)$ devrait à l'idéal être égal à 1 lorsque les vecteurs d'appartenance de $\mathbf{x}_k$ et $\mathbf{x}_l$ sont égaux. Remarquons d'abord que dans $\mathbb{M}_{fcn}$, les partitions ne sont pas forcément probabilistes. Ensuite, comme on ne dispose pas toujours en pratique des données $X$, on ne peut pas s'appuyer avec certitude sur les relations spatiales des individus. Enfin, à cause des limitations bien connues des partitions floues [Krishnapuram et Keller, 1993] qui n'expriment aucunement la typicalité des individus, il est délicat d'établir avec certitude que, par exemple,

deux individus $\mathbf{x}_k$ et $\mathbf{x}_l$ ayant le même vecteur d'appartenance ${}^t(0.5\ 0.5)$ appartiennent conjointement au même cluster, si bien qu'il est acceptable que $U_1(k,l) < 1$.

#### 3.2.1.2 Approches contingentielles

Les extensions fondées sur une approche contingentielle reposent sur le calcul de la matrice des contigences-paire via la matrice de contingence $N(U,V)$ croisant deux partitions, décrit page 28. Leur principal avantage réside dans leur faible complexité en temps. Mais, en s'appuyant sur les équations (2.21), (2.22), (2.23) et (2.25) issues de l'analyse combinatoire qui n'a pas lieu d'être dans les cas non-strictes, elle présentent en contrepartie des comportements indésirables, comme nous le montrons ci-après.

##### 3.2.1.2.a Anderson

Anderson et al. [2010] proposent d'utiliser directement l'équation (2.20) pour le calcul d'une matrice de contingence non-stricte $N_A$ croisant deux partitions non-strictes $U$ and $V$. Toutefois, afin de se prémunir d'éventuels comportements non souhaités lorsque la comparaison implique des partitions possibilistes dans $\mathbb{M}^{>}_{pcn}$, les auteurs proposent la normalisation suivante :

$$N_{A\star}(U,V) = \rho\, N_A(U,V), \tag{3.20}$$

où $\rho = \frac{n}{\sum_{i=1}^{c} n_{i\bullet}^{A}} = \frac{n}{\sum_{j=1}^{r} n_{\bullet j}^{A}}$ est un facteur de normalisation tel que $\rho = 1$ pour des partitions de $\mathbb{M}_{fcn}$.

À partir de cette matrice de contingence non-stricte $N_{A\star}$, les termes $m^{\top}_{\alpha\beta}$ $(\alpha,\beta \in \{0,1\})$ de la table des contingences-paire non-strictes $M_{A\star}(U,V)$ sont alors simplement calculés via les équations (2.21), (2.22), (2.23) et (2.25).

Puisque le produit est une norme triangulaire, Anderson et al. [2010] proposent aussi une autre formulation du calcul de la matrice $N_A$ pour laquelle le produit induit dans (2.20) est remplacé par n'importe quelle t-norme $\top$, amenant alors à la définition d'une matrice de contingence $N_A^{\top}$ dont les termes sont donnés par :

$$n_{ij}^{\top}(U,V) = \sum_{k=1}^{n} \top(u_{ik}, v_{jk}) \tag{3.21}$$

Dans ce cas général toutefois, le calcul du facteur de normalisation $\rho$, essentiel à la définition d'une matrice de contingence normalisée $N_{A\star}^{\top}$ et donc d'une table des contingences-paire $M_{A\star}^{\top}$ n'est pas explicité par les auteurs. Pourtant, un tel calcul n'est pas trivial. En effet, comme évoqué plus tôt, le facteur de normalisation est choisi tel que $\rho = 1$ lorsque les partitions considérées sont dans $\mathbb{M}_{fcn}$, de sorte que la

normalisation n'influe que sur la comparaison de partitions possibilistes. Or, tel que défini par ses auteurs, $\rho = \frac{n}{\sum_{i=1}^{c} n_{i\bullet}^{A}}$ ne présente cette propriété que dans l'unique cas où $N_A(U,V) = N_A^{\top_P}(U,V)$, pour laquelle $\sum_{i=1}^{c} n_{i\bullet}^{A} = n$ si $U$ et $V$ sont des partitions de $\mathbb{M}_{fcn}$. Pour toute autre t-norme, cette formulation de $\rho$ ne tient plus son rôle, si bien qu'en l'absence de toute précision à ce propos au sein de la littérature, on se restreindra dans la suite de ce mémoire à ne considérer uniquement que les mesures dérivées de $M_{A\star}^{\top_P}(U,V)$.

Cette normalisation est justifiée par Anderson et al. [2010] pour prévenir les cas où les mesures d'appariement dérivées pourraient prendre des valeurs négatives lorsque les partitions à comparer sont issues de $\mathbb{M}_{pcn}^{>}$, principalement à cause du terme $m_{00}$ de la table des contingences-paire qui peut dans ce cas lui-même être négatif. Ce comportement indésirable est toutefois à mettre en relation avec le choix fait par les auteurs de calculer ce terme à partir de l'équation (2.25) et non de l'équation (2.24), pourtant plus judicieuse puisque la relation telle que $\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij} = n$, sur laquelle repose (2.25) ne tient plus dans le cas du croisement de partitions possibilistes. Aussi, soulignons le fait que malgré la correction apportée par Anderson et al. [2010], les mesures dérivées de leur extension floue ne sont tout de même pas à l'abri de comportements singuliers, principalement à cause de l'utilisation de l'équation (2.21) utilisée pour le calcul du terme $m_{11}$ de la table des contingences-paire. En effet, puisque les termes de $N_{A\star}^{\top_P}$ peuvent être inférieurs à 1 lorsque $U$ et $V$ sont dans $\mathbb{M}_{pcn}$, rien n'empêche théoriquement le terme $n_{ij}^{\top_P}(n_{ij}^{\top_P} - 1)$ de cette équation de prendre des valeurs négatives, pouvant alors amener $m_{11}$ à lui-même être inférieur à 0. Bien que l'expérience a prouvé que ce comportement survient peu en pratique parce que les partitions alors considérées présentent bien plus d'individus que de clusters, nous proposons toutefois de corriger pour la suite cette proposition en écrêtant simplement à 0 les valeurs de $(n_{ij}^{\top_P} - 1)$ qui pourraient surgir lors du calcul du terme $m_{11}$ de $M_{A\star}^{\top_P}$. Remarquons que Hüllermeier et al. [2012], qui formulent une critique analogue à l'égard de la proposition d'Anderson, proposent une perspective de correction fondée sur la remarque du fait que dans le cas strict, l'équation (2.21) repose sur le coefficient binomial, tel que :

$$m_{11} = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}(n_{ij}-1) = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{r}\binom{n_{ij}}{2}. \tag{3.22}$$

Or, dans le cas où les termes $n_{ij}$ sont à valeur réelle, ce coefficient binomial se généralise habituellement à l'aide de la fonction Gamma, $\Gamma$, si bien que Hüllermeier et al. [2012] proposent ainsi une autre extension du terme $m_{11}$ :

$$m_{11} = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{\Gamma(n_{ij}+1)}{\Gamma(n_{ij})} \tag{3.23}$$

Enfin, comme le mentionnent ses auteurs, les mesures dérivées de cette approche perdent la propriété de réflexivité *(II-1)*, si bien qu'aucune d'entre elles ne préservent le cas échéant leur qualité de métrique sur $\mathbb{M}_{scn}$.

**Exemple 20.** *Considérons les deux partitions floues $U_f$ et $V_f$ (page 42). Leur matrice des contingence normalisée, calculée selon l'approche proposée par Anderson et al. avec la t-norme produit $\top_P$ vaut :*

$$N_{A\star}(U_f, V_f)^{\top_P} = \begin{pmatrix} 1.170 & 1.140 & 0.590 \\ 0.430 & 0.560 & 1.110 \end{pmatrix}.$$

*Les matrices des contingences-paire induite avant et après correction valent alors :*

$$M_{A\star}^{\top_P}(U_f, V_f) = \begin{pmatrix} -0.126 & 4.036 \\ 1.796 & 4.294 \end{pmatrix} \quad et \quad M_{\widetilde{A\star}}^{\top_P}(U_f, V_f) = \begin{pmatrix} 0.240 & 4.036 \\ 1.796 & 4.294 \end{pmatrix}.$$

*Les valeurs des indices corrigés de Rand $RI_{\widetilde{A\star}}^{\top_P}$ et de Jaccard $JI_{\widetilde{A\star}}^{\top_P}$ obtenus en croisant toutes les partitions exemples $U_f$, $V_f$ et $W_f$ sont données à la Table 3.6. On remarque que les valeurs de $RI_{\widetilde{A\star}}^{\top_P}$ sont légèrement inférieures à celles obtenues avec la proposition de Campello et restent cohérentes sans pour autant témoigner d'une dynamique importante (propriété (II-4)). On constate que ces indices perdent ici aussi leur propriété de réflexivité (I-1). Par ailleurs, on note que les valeurs de $JI_{\widetilde{A\star}}^{\top_P}$ sont très faibles. Cela s'explique aisément lorsque l'on considère la matrice des contingences-paire corrigée $M_{\widetilde{A\star}}^{\top_P}$ pour laquelle le terme $m_{11}$, qui est le seul à intervenir au numérateur de l'indice de Jaccard (2.11), est très faible par rapport aux termes $m_{01}$ et $m_{10}$. Par ailleurs, on remarque encore que $JI_{\widetilde{A\star}}^{\top_P}(W_f, W_f) < JI_{\widetilde{A\star}}^{\top_P}(W_f, U_f)$, pour les raisons évoquées à l'Exemple 19.*

| $RI_{\widetilde{A\star}}^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.453 | 0.437 | 0.450 |
| $V_f$ | 0.437 | 0.542 | 0.577 |
| $W_f$ | 0.450 | 0.577 | 0.627 |

| $JI_{\widetilde{A\star}}^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.187 | 0.040 | 0.011 |
| $V_f$ | 0.040 | 0.001 | 0.000 |
| $W_f$ | 0.011 | 0.000 | 0.000 |

TABLE 3.6 – $RI_{\widetilde{A\star}}^{\top_P}$ (à gauche) et $JI_{\widetilde{A\star}}^{\top_P}$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

#### 3.2.1.2.b Ceccarelli et Maratea

Ceccarelli et Maratea [2008] proposent une autre définition de la matrice de contingence croisant deux partitions non-strictes, notée $N_{CM}^{\gamma}$, dérivée encore de l'équation (2.20) :

$$n_{ij}^{\gamma}(U, V) = \sum_{k=1}^{n} (u_{ik} + v_{jk})^{\gamma} \tag{3.24}$$

où $\gamma \in [1, +\infty[$ est choisi pour améliorer la robustesse des mesures d'appariement dérivées en accentuant l'influence des degrés d'appartenance les plus forts et en amenuisant simultanément l'influence des degrés les plus faibles. Selon cette approche, les termes de la table des contingences-paire sont obtenus encore une fois à partir des équations (2.21), (2.22), (2.23) et (2.24) pour lesquelles les termes $n_{ij}$ sont remplacés par ceux donnés par l'équation (3.24). On note $M^{\gamma}_{CM}(U, V)$ la table des contingences-paire resultante, à partir de laquelle toutes les mesures d'appariement décrites au Chapitre 2 peuvent être étendues aux cas flous et possibilistes.

Encore une fois, les mesures dérivées de cette approche ne sont pas réflexives selon la propriété *(II-1)* pour toute partition stricte ou non de $\mathbb{M}_{pcn}$. De plus, à cause de la somme utilisée pour l'agrégation des degrés d'appartenance dans l'équation (3.24), elles ne se ramènent pas à leur pendant strict lorsque les partitions comparées sont dans $\mathbb{M}_{hcn}$, contrairement aux autres extensions revues dans ce chapitre. Ce n'est donc pas une généralisation au sens où nous l'entendons. Pour cette raison nous exclurons cette proposition des études expérimentales comparant nos propositions à celles de la littérature (Chapitre 4).

**Exemple 21.** *Considérons les deux partitions floues $U_f$ et $V_f$ (page 42). Les matrices de contingences et de contingences-paire calculées selon l'approche de Ceccarelli et Maratea avec $\gamma = 2$ valent respectivement :*

$$N^2_{CM}(U_f, V_f) = \begin{pmatrix} 5.230 & 5.260 & 4.260 \\ 2.950 & 3.300 & 4.500 \end{pmatrix} \quad et \quad M^2_{CM}(U_f, V_f) = \begin{pmatrix} 43.755 & 110.057 \\ 51.956 & 106.606 \end{pmatrix}.$$

*Les valeurs des indices de Rand $RI^2_{CM}(U_f, V_f)$ et de Jaccard $JI^2_{CM}(U_f, V_f)$ induites par cette dernière sont données, ainsi que toutes celles obtenues en croisant toutes les partitions-exemple $U_f$, $V_f$ et $W_f$, à la Table 3.7. Les valeurs des deux indices sont encore une fois comparables à celles obtenues pour les extensions décrites précédemment et présentent une faible dynamique. On constate aussi que $JI^2_{CM}(W_h, W_h) < JI^2_{CM}(W_h, U_h)$, illustrant la non réflexivité des mesures dérivées de cette approche.*

| $RI^2_{CM}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.488 | 0.481 | 0.479 |
| $V_f$ | 0.481 | 0.545 | 0.571 |
| $W_f$ | 0.479 | 0.571 | 0.610 |

| $JI^2_{CM}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.312 | 0.213 | 0.162 |
| $V_f$ | 0.213 | 0.150 | 0.112 |
| $W_f$ | 0.162 | 0.112 | 0.082 |

TABLE 3.7 – $RI^2_{CM}$ (à gauche) et $JI^2_{CM}$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

#### 3.2.1.3 Approches coïncidentielles

##### 3.2.1.3.a Borgelt

Afin d'étendre la notion de matrice de coïncidence $\Psi_U$ définie par (2.27) au partitions non-strictes Borgelt [2006] propose d'utiliser encore une fois des normes triangulaires. Soit $\Psi_U^{\top}$ la matrice de coïncidence non-stricte d'une partition floue $U \in \mathbb{M}_{fcn}$, son terme général est donné, pour n'importe quelle t-norme $\top$, par :

$$\psi_{U,kl}^{\top} = \sum_{i=1}^{c} \top(u_{ik}, u_{il}). \tag{3.25}$$

À partir de l'équation (3.25) est calculée la table des contingences-paire $M_B^{\top}(U,V)$ en étendant l'équation (2.28) au domaine non-stricte. Ainsi, les quatre termes $m_{\alpha\beta}^{\top}$ $(\alpha,\beta \in \{0,1\})$ de $M(U,V)$ deviennent, pour n'importe quel couple de t-normes $(\top, \perp)$ :

$$m_{\alpha\beta}^{\top}(\Psi_U^{\top}, \Psi_V^{\top}) = \sum_{k=2}^{n}\sum_{l=1}^{k-1} \top\big((1-\alpha) + (2\alpha - 1)\ \psi_{U,kl}^{\top}, (1-\beta) + (2\beta - 1)\ \psi_{V,kl}^{\top}\big). \tag{3.26}$$

Cette table de contingence-paires permet bien évidemment de dériver toutes les mesures d'appariement décrites au Chapitre 2, notamment les indices de Rand $RI_B^{\top}$ et de Jaccard $JI_B^{\top}$. Encore une fois, les mesures dérivées de cette approche ne sont pas réflexives selon *(II-1)* pour les partitions de $\mathbb{M}_{scn}$, si bien qu'aucune d'entre elle ne peut être une métrique, ou même une similarité. Par ailleurs, comme le remarquent Hüllermeier et al. [2012], cette approche est strictement équivalente, lorsqu'elle est calculée avec la t-norme produit $\top_P$, à celle proposée par Campello [2007] lorsque cette dernière est elle même calculée avec le couple de t-norme/ t-conorme $(\top_P, \perp_Ł)$.

**Exemple 22.** *Les matrices de coïncidences des partitions floues $U_f$ et $V_f$ (page 42), calculées avec la t-norme produit $\top_P$ valent :*

$$\Psi_{U_f}^{\top_P} = \begin{pmatrix} 0.820 & 0.260 & 0.740 & 0.660 & 0.340 \\ 0.260 & 0.680 & 0.320 & 0.380 & 0.620 \\ 0.740 & 0.320 & 0.680 & 0.620 & 0.380 \\ 0.660 & 0.380 & 0.620 & 0.580 & 0.420 \\ 0.340 & 0.620 & 0.380 & 0.420 & 0.580 \end{pmatrix}$$

*et*

$$\Psi_{V_f}^{\top_P} = \begin{pmatrix} 0.660 & 0.170 & 0.240 & 0.310 & 0.240 \\ 0.170 & 0.540 & 0.230 & 0.220 & 0.530 \\ 0.240 & 0.230 & 0.540 & 0.490 & 0.180 \\ 0.310 & 0.220 & 0.490 & 0.460 & 0.190 \\ 0.240 & 0.530 & 0.180 & 0.190 & 0.540 \end{pmatrix}.$$

*La matrice des contingences-paire induite par (3.26) vaut alors :*

$$M_B^{\top_P}(U_f, V_f) = \begin{pmatrix} 1.446 & 3.294 \\ 1.354 & 3.906 \end{pmatrix}.$$

*Les valeurs des indices de Rand $RI_B^{\top_P}(U_f, V_f)$ et de Jaccard $JI_B^{\top_P}(U_f, V_f)$ calculés à partir de cette dernière sont données, ainsi que toutes celles obtenues en croisant toutes les partitions-exemple $U_f$, $V_f$ et $W_f$, à la Table 3.8. Comme pour toutes les autres propositions présentées plus tôt, la dynamique des indices dérivés de la proposition de Borgelt n'est pas exceptionnelle, notamment pour l'indice de Rand. De plus, on remarque à nouveau qu'à cause de la perte de la propriété de réflexivité certains comportements contre intuitifs tels que $JI_B^{\top_P}(W_h, W_h) < JI_B^{\top_P}(W_h, U_h)$ peuvent être constatés.*

| $RI_B^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.553 | 0.535 | 0.537 |
| $V_f$ | 0.535 | 0.626 | 0.653 |
| $W_f$ | 0.537 | 0.653 | 0.698 |

| $JI_B^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.359 | 0.237 | 0.189 |
| $V_f$ | 0.237 | 0.199 | 0.164 |
| $W_f$ | 0.189 | 0.164 | 0.149 |

TABLE 3.8 – $RI_B^{\top_P}$ (à gauche) et $JI_B^{\top_P}$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

**Théorème 2.** *Les formules (2.37), (2.38) et (2.39) de Marcotorchino [1984] reliant les termes de la matrice de contingence non-stricte (page 56) et les termes des matrices de coïncidence floues (3.25) restent vérifiées pour la t-norme produit $\top_P$ pour deux partitions floues $U$ et $V$ de $\mathbb{M}_{fcn}$.*

*Démonstration.* On a :

$$\begin{aligned} n_{i\bullet} &= \sum_{j=1}^{r} n_{ij} \\ &= \sum_{j=1}^{r} \sum_{k=1}^{n} u_{ik}\, v_{jk} \\ &= \sum_{k=1}^{n} u_{ik} \sum_{j=1}^{r} v_{jk} \\ &= \sum_{k=1}^{n} u_{ik} \end{aligned}$$

puisque $V \in \mathbb{M}_{fcn}$.
Alors :

$$
\begin{aligned}
n_{i\bullet}^2 &= \left( \sum_{k=1}^{n} u_{ik} \right)^2 \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} u_{ik}\, u_{il}
\end{aligned}
$$

et finalement :

$$
\begin{aligned}
\sum_{i=1}^{c} n_{i\bullet}^2 &= \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{l=1}^{n} u_{ik}\, u_{il} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{i=1}^{c} u_{ik}\, u_{il} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \psi_{U,kl}^{\top_p}
\end{aligned}
$$

Ce qui prouve (2.37) et (2.38), la preuve de cette dernière formule étant analogue. Enfin, pour (2.39), on a :

$$
\begin{aligned}
\sum_{i=1}^{c} \sum_{j=1}^{r} n_{ij}^2 &= \sum_{i=1}^{c} \sum_{j=1}^{r} n_{ij} \sum_{l=1}^{n} u_{il}\, v_{jl} \\
&= \sum_{i=1}^{c} \sum_{j=1}^{r} \sum_{k=1}^{n} u_{ik}\, v_{jk} \sum_{l=1}^{n} u_{il}\, v_{jl} \\
&= \sum_{i=1}^{c} \sum_{j=1}^{r} \sum_{k=1}^{n} \sum_{l=1}^{n} u_{ik}\, v_{jk}\; u_{il}\, v_{jl} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{i=1}^{c} u_{ik}\, u_{il} \sum_{j=1}^{r} v_{jk}\, v_{jl} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{i=1}^{c} u_{ik}\, u_{il}\; \psi_{V,kl}^{\top_p} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \psi_{U,kl}^{\top_p}\, \psi_{V,kl}^{\top_p}
\end{aligned}
$$

Ce qui clôt la preuve. □

Cette approche coïncidentielle a été très récemment vivement critiquée, notamment par Hüllermeier et Rifqi [2009] qui en formulent une critique identique à celle discutée page 53 mais aussi par Campello [2010] qui lui reproche son manquement à la propriété d'identité des indiscernables *(I-2)*. Par construction, il est en effet possible, même si c'est loin d'être trivial, de trouver, dans $\mathbb{M}_{scn}$, deux partitions différentes qui

ont exactement la même matrice de coïncidence. Dans un tel cas de figure, en pratique peu fréquent, les deux partitions concernées seraient alors considérées comme étant en totale concordance pour toute mesure dérivée.

#### 3.2.1.3.b Brouwer

Brouwer [2009a] soulève un problème intéressant dans l'extension des matrices de coïncidence telles que définies par Borgelt (page 60). En effet, à cause de l'utilisation d'une t-norme dans (3.25), les termes diagonaux $\psi^{\top}_{U,kk}$ d'une telle matrice calculée pour une partition floue $U \in \mathbb{M}_{fcn}$, représentant le degré selon lequel chaque individu $\mathbf{x}_k$ est dans le même cluster que lui-même ne sont généralement pas égaux à 1, et peuvent même en être très inférieurs. Afin de résoudre ce comportement contre-intuitif, Brouwer, qui ne considère que le produit $\top_P$, propose de remplacer les termes $\psi^{\top}_{U,kl}$ du calcul de $\Psi^{\top}_U$ par :

$$\phi^{\top_P}_{U,kl} = \frac{\psi^{\top_P}_{U,kl}}{\sqrt{\psi^{\top_P}_{U,kk}}\sqrt{\psi^{\top_P}_{U,ll}}} \tag{3.27}$$

de telle sorte de transformer $\Psi^{\top_P}_U$ en une nouvelle matrice de coïncidence $\Phi^{\top_P}_U$ dont les termes diagonaux sont égaux à 1. Cette normalisation présente entre autres la propriété d'être invariante en échelle $\forall \alpha \in \mathbb{R}^+$,

$$\Phi^{\top_P}_{\alpha U} = \Phi^{\top_P}_U. \tag{3.28}$$

En pratique, elle améliore la dynamique (propriété (II-4)) des mesures dérivées, comme le montrent l'Exemple 23 et les expérimentations menées au Chapitre 4. Nous proposerons de généraliser cette normalisation à toute norme triangulaire au Chapitre 4, page 85. C'est la raison des notations $M^{\top_P}_Q$, $RI^{\top}_Q$ et $JI^{\top_P}_Q$ ci-après.

**Exemple 23.** *Les matrices de coïncidence normalisées des partitions floues $U_f$ et $V_f$ (page 42), calculées avec la t-norme produit $\top_P$, à comparer aux matrices non normalisées de l'Exemple 22, valent respectivement :*

$$\Phi^{\top_P}_{U_f} = \begin{pmatrix} 1.000 & 0.348 & 0.991 & 0.957 & 0.493 \\ 0.348 & 1.000 & 0.471 & 0.605 & 0.987 \\ 0.991 & 0.471 & 1.000 & 0.987 & 0.605 \\ 0.957 & 0.605 & 0.987 & 1.000 & 0.724 \\ 0.493 & 0.987 & 0.605 & 0.724 & 1.000 \end{pmatrix}$$

*et*

$$\Phi^{\top_P}_{V_f} = \begin{pmatrix} 1.000 & 0.285 & 0.402 & 0.563 & 0.402 \\ 0.285 & 1.000 & 0.426 & 0.441 & 0.981 \\ 0.402 & 0.426 & 1.000 & 0.983 & 0.333 \\ 0.563 & 0.441 & 0.983 & 1.000 & 0.381 \\ 0.402 & 0.981 & 0.333 & 0.381 & 1.000 \end{pmatrix}.$$

*La matrice des coincidences-paire résultante, à comparer à* $M_B^{\top_P}(U_f, V_f)$ *(donnée à l'Exemple 22, page 60), vaut alors :*

$$M_Q^{\top_P}(U_f, V_f) = \begin{pmatrix} 4.119 & 3.050 \\ 1.079 & 1.752 \end{pmatrix}.$$

*Les valeurs des indices de Rand et de Jaccard calculées, pour tous les couples de partitions-exemple* $U_f$, $V_f$ *et* $W_f$, *à partir des matrices de coïncidences normalisées sont donnés à la Table 3.9 pour la t-norme produit* $\top_P$. *En comparant les matrices de coïncidences normalisées selon cette proposition avec celles, non-normalisées, de l'Exemple 22, on constate qu'en plus des termes diagonaux ramenés à 1, les éléments non diagonaux de ces matrices sont aussi renforcés, mais non linéairement. Ainsi, le degré de co-appartenance normalisé* $\phi_{U_f}^{\top_P}(2,1)$ *de* $\{\mathbf{x}_2, \mathbf{x}_1\}$ *est proportionnellement plus grand que celui* $\phi_{U_f}^{\top_P}(2,4)$ *de* $\{\mathbf{x}_2, \mathbf{x}_1\}$ *par rapport à ceux de la matrice de coïncidence non normalisée* $\Psi_{U_h}^{\top_P}$. *Il découle de cela que les indices dérivés de la proposition de Brouwer présentent une meilleure dynamique que les propositions précédentes. Toutefois, elle présente les mêmes inconvénients en ne préservant pas la propriété de réflexivité (I-1) des indices dérivés.*

| $RI_Q^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.704 | 0.587 | 0.547 |
| $V_f$ | 0.587 | 0.617 | 0.599 |
| $W_f$ | 0.547 | 0.599 | 0.613 |

| $JI_Q^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 0.658 | 0.499 | 0.436 |
| $V_f$ | 0.499 | 0.462 | 0.409 |
| $W_f$ | 0.436 | 0.409 | 0.385 |

TABLE 3.9 – $RI_Q^{\top_P}$ (à gauche) et $JI_Q^{\top_P}$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

### 3.2.2 Approche orientée clusters : distance de transfert non-stricte

Campello [2010] propose d'étendre et de généraliser aux cas non-stricts la distance de transfert $TD$ de Charon et al. [2006] décrite page 33. L'auteur définit la *distance de transfert floue* $FTD(U, V)$ (*Fuzzy Transfert Distance*) entre deux partitions non strictes $U$ et $V$ de $\mathbb{M}_{scn}$, non plus comme le nombre d'éléments à transférer des clusters de $U$ pour qu'ils coïncident, à une permutation près, avec ceux de $V$, mais comme la quantité minimum de degrés d'appartenance qui doit être ajoutée et/ou retirée aux colonnes de $U$ pour les rendre égales à celles de $V$. Autrement dit, chaque arête $(i, j)$ du graphe biparti construit pour le calcul de cette distance est valuée par :

$$w_{ij} = \sum_{k=1}^{n} \mid u_{ik} - v_{jk} \mid . \tag{3.29}$$

À l'instar de son pendant strict, la distance de transfert floue peut être calculée efficacement par la méthode Hongroise. Elle est par ailleurs bornée supérieurement, si bien que son auteur préconise la normalisation suivante :

$$FTD_N(U,V) = \frac{FTD(U,V)}{n \times max(c,r)} \tag{3.30}$$

de sorte que $FTD_N(U,V) \in [0,1]$. On peut alors dériver l'indice $I_{FTD_N}(U,V) = 1 - FTD_N$. Notons finalement que l'idée de cette extension au cas non-strict de la distance de transfert fut préalablement évoquée dans le même contexte par Gusfield [2002].

**Exemple 24.** *Les valeurs de $I_{FTD_N}$, calculées pour les partitions $U_f$, $V_f$ et $W_f$ (page 42), sont données par la Table 3.10. On remarque que cet indice est réflexif (propriété (I-1)), contrairement aux propositions précédemment revues. Par ailleurs, sa dynamique est légèrement plus forte que celle dont témoignent les précédentes. Enfin, on constate que $I_{FTD_N}(U_f,U_f) > I_{FTD_N}(U_f,V_f) > I_{FTD_N}(U_f,W_f)$, tel que l'on pourrait intuitivement s'y attendre.*

| $I_{FTD_N}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 1.000 | 0.787 | 0.780 |
| $V_f$ | 0.787 | 1.000 | 0.900 |
| $W_f$ | 0.780 | 0.900 | 1.000 |

TABLE 3.10 – $I_{FTD_N}$ pour les partitions $U_f$, $V_f$ et $W_f$.

## 3.3 Approches natives

### 3.3.1 Approche orientée individus : indices de Huellermeier et al.

L'approche proposée par Hüllermeier et Rifqi [2009] a cela de particulier qu'au contraire des propositions précédemment revues dans cette section, elle ne permet à l'origine d'étendre au domaine non-strict qu'uniquement l'indice de Rand. Sa construction, proposée pour pallier les critiques formulées à l'encontre des mesures dérivées de l'extension proposée par Campello [2007], page 53, repose par ailleurs sur une approche originale, dite géométrique car fondée sur la mesure d'une distance entre les vecteurs d'appartenances de chacune des partitions à comparer.

Pour chaque partition floue $U$ et $V$ , on définit ainsi une matrice de similarité notée $\Psi_U^{HR}$ (respectivement $\Psi_V^{HR}$) par analogie avec les matrices de coïncidence [4], dont le

4. Les auteurs utilisent la notation $E_U(\mathbf{x}_k, \mathbf{x}_l)$ symbolisant une mesure d'équivalence de la paire $(k,l)$ du point de vue de $U$.

terme général est calculé par :

$$\psi^{HR}_{U,kl} = 1 - d(\mathbf{u}_k - \mathbf{u}_l) \tag{3.31}$$

ou $d$ est une métrique telle que $\psi^{HR}_{U,kl} \in [0,1]$ et bien évidemment telle que $\psi^{HR}_{U,kk} = 1$. Les auteurs préconisent d'utiliser la distance de Manhattan normalisée (aussi appelée norme $L_1$), définie par :

$$d_1(\mathbf{u}_k, \mathbf{u}_l) = \sum_{i=1}^{c} \frac{\mid u_{ik} - u_{il} \mid}{2}. \tag{3.32}$$

**Théorème 3.** *La matrice $\Psi^{HR}_U$ calculée avec la norme $L_1$ est la matrice de coïncidence $\Psi^{\top}_U$ calculée avec la t-norme $\top_M$ définie par Borgelt [2006] et présentée page 60.*

*Démonstration.* Comme $min(a,b) = \frac{a+b-|a-b|}{2}$, $\forall a, b$, on a :

$$\begin{aligned}
\psi^{\top_M}_{U,kl} &= \sum_{i=1}^{c} min(u_{ik}, u_{il}) \\
&= \sum_{i=1}^{c} \frac{u_{ik} + u_{il} - \mid u_{ik} - u_{il} \mid}{2} \\
&= \sum_{i=1}^{c} \frac{u_{ik} + u_{il}}{2} - \sum_{i=1}^{c} \frac{\mid u_{ik} - u_{il} \mid}{2} \\
&= 1 - \sum_{i=1}^{c} \frac{\mid u_{ik} - u_{il} \mid}{2} \\
&= \psi^{HR}_{U,kl}
\end{aligned}$$

□

Nous proposons de prendre n'importe quelle distance de Minkowski normalisée :

$$d_p(\mathbf{u}_k, \mathbf{u}_l) = \frac{(\sum_{i=1}^{c} \mid u_{ik} - u_{il} \mid^p)^{\frac{1}{p}}}{2^{\frac{1}{p}}}. \tag{3.33}$$

L'indice de Rand de Huellermeier et Rifqi est alors obtenu en agrégeant les deux matrices de similarité calculées pour $U$ et $V$ comme suit :

$$RI^p_{HR}(U,V) = \frac{1}{q} \sum_{k=2}^{n} \sum_{l=1}^{k-1} 1 - \mid \psi^{HR}_{U,kl} - \psi^{HR}_{V,kl} \mid \tag{3.34}$$

où $p$ est le paramètre de la mesure de Minkowski définie par (3.33) et où $q = \frac{n(n-1)}{2}$ est le nombre de paires différentes d'individus $\{\mathbf{x}_k, \mathbf{x}_l\}$. Comme le soulignent ses auteurs, la mesure de distance dérivée $D^p_{HR}(U,V) = 1 - RI^p_{HR}(U,V)$ est une pseudométrique et

satisfait donc l'inégalité triangulaire (I-4) si bien que $RI_{HR}$ se révèle être une mesure relative de choix. Par ailleurs, il s'agit d'une généralisation de l'indice de Rand, si bien que dans le cas de la comparaison de partitions strictes, $RI^p_{HR}$ est tout à fait égal à $RI$.

**Théorème 4.** *L'indice de Rand généralisé de Hüllermeier et Rifqi [2009] calculé avec la norme $L_1$ peut être exprimé à partir des termes de la matrice de contingence-paires $M^{\top_M}$ définie par Borgelt [2006].*

*Démonstration.* Comme $min(a,b) + min(1-a, 1-b) = 1 - \mid a - b \mid$, on a :

$$\begin{aligned} RI^1_{HR}(U,V) &= \frac{1}{q}\sum_{k=2}^{n}\sum_{l=1}^{k-1} 1 - \mid \psi^{HR}_{U,kl} - \psi^{HR}_{V,kl} \mid \\ &= \frac{1}{q}\sum_{k=2}^{n}\sum_{l=1}^{k-1} min(\psi_{U,kl}, \psi_{V,kl}) + min(1-\psi_{U,kl}, 1-\psi_{V,kl}) \\ &= \frac{m_{11}^{\top_M} + m_{00}^{\top_M}}{q} \quad \text{par (3.26).} \end{aligned}$$

□

Dans [Hüllermeier et al., 2012], les mêmes auteurs présentent une extension de leur proposition à toutes les mesures dites "par appariement", en redéfinissant les termes $m_{11}$, $m_{10}$, $m_{01}$ et $m_{00}$ de la matrice des contingences-paires. Cette généralisation se formule comme suit :

$$m^{HR}_{11} = \sum_{k=2}^{n}\sum_{l=1}^{k-1} \top\left(1 - \mid \psi^{HR}_{U,kl} - \psi^{HR}_{V,kl} \mid,\ \top(\psi^{HR}_{U,kl}, \psi^{HR}_{V,kl})\right) \tag{3.35}$$

$$m^{HR}_{10} = \sum_{k=2}^{n}\sum_{l=1}^{k-1} max(\psi^{HR}_{U,kl} - \psi^{HR}_{V,kl},\ 0) \tag{3.36}$$

$$m^{HR}_{01} = \sum_{k=2}^{n}\sum_{l=1}^{k-1} max(\psi^{HR}_{V,kl} - \psi^{HR}_{U,kl},\ 0) \tag{3.37}$$

$$m^{HR}_{00} = \sum_{k=2}^{n}\sum_{l=1}^{k-1} \top\left(1 - \mid \psi^{HR}_{U,kl} - \psi^{HR}_{V,kl} \mid,\ \bot(1-\psi^{HR}_{U,kl}, 1-\psi^{HR}_{V,kl})\right), \tag{3.38}$$

où $\top$ et $\bot$ sont respectivement une t-norme et sa t-conorme duale. Toutefois, pour que l'égalité :

$$m_{11} + m_{00} = 1 - \mid \psi^{HR}_{U,kl} - \psi^{HR}_{V,kl} \mid \tag{3.39}$$

soit vérifiée de telle sorte que (3.34) puisse être dérivé de (3.35), (3.36), (3.37) et (3.38),

seul le couple produit $(\top_P, \perp_P)$ peut être utilisé, si bien que l'on a finalement :

$$m_{11}^{HR} = \sum_{k=2}^{n}\sum_{l=1}^{k-1}\left(1-\mid \psi_{U,kl}^{HR} - \psi_{V,kl}^{HR} \mid\right) \times \left(\psi_{U,kl}^{HR} \times \psi_{V,kl}^{HR}\right) \tag{3.40}$$

$$m_{00}^{HR} = \sum_{k=2}^{n}\sum_{l=1}^{k-1}\left(1-\mid \psi_{U,kl}^{HR} - \psi_{V,kl}^{HR} \mid\right) \times \left(1 - \psi_{U,kl}^{HR} \times \psi_{V,kl}^{HR}\right). \tag{3.41}$$

Remarquons alors que cette proposition d'extension présente de fortes analogies avec celle faite par Quéré et Frélicot [2011b], détaillée au chapitre 4, section 4.5.

**Exemple 25.** *Les valeurs de $RI_{HR}^1$ et de $JI_{HR}^1$ pour les partitions-exemple $U_f$, $V_f$ et $W_f$ (page 42) sont données à la Table 3.11. Cette proposition préservant les propriétés des indices qu'elle généralise, il est normal de constater que $RI_{HR}^1$ et $JI_{HR}^1$ sont réflexifs et tels que $I(U_f, U_f) > I(U_f, V_f) > I(U_f, W_f)$. Aussi, on observe que la dynamique de ces indices est bien meilleure que celle présentée par les indices déjà revus.*

| $RI_{HR}^1$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 1.000 | 0.880 | 0.790 |
| $V_f$ | 0.880 | 1.000 | 0.910 |
| $W_f$ | 0.790 | 0.910 | 1.000 |

| $JI_{HR}^1$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 1.000 | 0.717 | 0.528 |
| $V_f$ | 0.717 | 1.000 | 0.715 |
| $W_f$ | 0.528 | 0.715 | 1.000 |

TABLE 3.11 – $RI_{HR}^1$ (à gauche) et $JI_{HR}^1$ (à droite) pour les partitions $U_f$, $V_f$ et $W_f$.

### 3.3.2 Approches orientées clusters

#### 3.3.2.1 Mesures de compatibilité

Les mesures de comparaison décrites dans cette section sont fondées sur des mesures de compatibilité entre clusters stricts et flous, telles que décrites respectivement au Chapitre 2, page 32, et au présent Chapitre 3, page 52. Pour chaque proposition, nous donnerons l'expression de la mesure de comparaison telle que donnée par le ou les auteurs, pour ensuite la reformuler afin d'expliciter et de caractériser la mesure de compatibilité utilisée.

##### 3.3.2.1.a Beringer et Hüllermeier

Beringer et Hüllermeier [2007] proposent un indice symétrique pour la comparaison de deux partitions $U$ et $V$ de $\mathbb{M}_{pcn}$ fondé sur une mesure de similarité entre

les clusters $U^i$ et $V^j$ de chaque partition, définie par :

$$s(U^i, V^j) = \frac{|U^i \cap V^j|}{|U^i \cup V^j|} \tag{3.42}$$

$$= \frac{\sum_{k=1}^{n} min(u_{ik}, v_{jk})}{\sum_{k=1}^{n} max(u_{ik}, v_{jk})} \tag{3.43}$$

où (3.42) est une mesure de similarité entre ensembles bien connue, que l'on attribue généralement à Jaccard [1901]. Il s'agit d'ailleurs d'une mesure de ressemblance (Définition 21, page 52), avec $F(a, b, c) = \frac{a}{a+b+c}$.

À partir de cette ressemblance, on définit la mesure de concordance, asymétrique, de la partition $V$ par rapport à la partition $U$ par :

$$s(U, V) = \top_{i=1}^{c} \bot_{j=1}^{r} s(U^i, V^j). \tag{3.44}$$

Enfin, l'indice proposé se formule ainsi :

$$S_{BH}^{\top}(U, V) = \top\big(s(U, V), s(V, U)\big) \tag{3.45}$$

Il est réflexif (propriété *(I-1)*) et présente une bonne dynamique (propriété *(II-4)*), comme l'illustre l'Exemple 26.

**Exemple 26.** *Les valeurs de l'indice $S_{BH}^{\top}$, utilisé pour comparer les partitions $U_f$, $V_f$ et $W_f$ (page 42), sont données à la Table 3.12 pour la t-norme produit $\top_P$. La relation d'ordre que l'on peut déduire de ses valeurs coïncide avec celle que l'on perçoit intuitivement, à savoir que $U_f$ est plus proche de $V_f$ que de $W_h$.*

| $S_{BH}^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 1.000 | 0.328 | 0.108 |
| $V_f$ | 0.328 | 1.000 | 0.423 |
| $W_f$ | 0.108 | 0.423 | 1.000 |

TABLE 3.12 – $S_{BH}^{\top_P}$ pour les partitions $U_f$, $V_f$ et $W_f$.

Afin de prendre en compte dans la mesure la taille des clusters de chaque partition, les auteurs proposent aussi de remplacer (3.44) dans (3.45) par sa généralisation suivante, fondée sur une t-norme pondérée [Kaymak et van Nauta Lemke, 1998] :

$$s_w(U, V) = \top_{i=1}^{c} \mathcal{I}\Big(w_i, \bot_{j=1}^{r} s(U^i, V^i)\Big) \tag{3.46}$$

où $\mathcal{I}$ est une implication floue telle que définie à la section 3.1.2 et où $w_i = \frac{\sum_{k=1}^{n} u_{ik}}{n}$ est le poids de la classe $U^i$, calculé selon sa cardinalité floue. Ainsi, plus un cluster présente de forts degrés d'appartenance, plus il contribue à la mesure de concordance $s_w(U, V)$.

#### 3.3.2.1.b Runkler

Runkler [2010] propose un indice de comparaison tout à fait analogue, qui ne diffère que par l'utilisation d'une t-conorme en lieu et place de la t-norme dans (3.45) :

$$S_R^{\top}(U,V) = \perp\big(s(U,V), s(V,U)\big). \tag{3.47}$$

où $s(U,V)$ est donné par (3.44). L'auteur remarque d'ailleurs lui-même la proximité entre les deux propositions et note que celle de Beringer et Hüllermeier [2007] est plus pessimiste que la sienne, par construction ($\top$ à la place de $\perp$), de sorte que $S_{BH}^{\top}(U,V)$ présente une meilleure dynamique que $S_R^{\top}(U,V)$, comme l'illustre bien l'Exemple 27. C'est la raison pour laquelle nous ne retiendrons que l'indice de Beringer et Hüllermeier [2007] pour les expérimentations présentées au Chapitre 4.

**Exemple 27.** *La valeurs de l'indice $S_R^{\top}$, obtenues pour les partitions $U_f$, $V_f$ et $W_f$ (page 42), sont donnés à la Table 3.13 pour la t-norme produit $\top_P$.*

| $S_R^{\top_P}$ | $U_f$ | $V_f$ | $W_f$ |
|---|---|---|---|
| $U_f$ | 1.000 | 0.871 | 0.776 |
| $V_f$ | 0.871 | 1.000 | 0.910 |
| $W_f$ | 0.776 | 0.910 | 1.000 |

TABLE 3.13 – $S_R^{\top_P}$ pour les partitions $U_f$, $V_f$ et $W_f$.

#### 3.3.2.1.c Bodjanova

Bodjanova [1999] propose une mesure de concordance entre deux partitions floues $U$ et $V$ de même taille ($c = r$), fondée sur la comparaison de leur approximation stricte respective.

**Définition 22.** *La coupe de niveau $\alpha \in [0,1]$, ou $\alpha$-coupe, $U_\alpha^i$ d'un ensemble flou $U^i$ est une approximation stricte de ce dernier, dont les termes $(u_\alpha)_{ik}$ se définissent comme suit :*

$$(u_\alpha)_{ik} = \begin{cases} 1 & \text{si et seulement si } u_{ik} \geq \alpha \\ 0 & \text{sinon} \end{cases} \tag{3.48}$$

*La coupe $\alpha$ d'une partition floue $U$ se constitue des $\alpha$-coupes $U_\alpha^i$ des clusters de $U$ et est notée $U_\alpha$.*

Deux partitions $U$ et $V$ de même taille sont ainsi dites $\alpha$-équivalentes par l'auteur si et seulement si $U_\alpha \equiv V_\alpha$. Une telle relation est notée $U =_\alpha V$. Une mesure de

dissimilarité entre $U_\alpha$ et $V_\alpha$ est alors introduite par :

$$D_\alpha(U,V) = \frac{1}{nc}\sum_{i=1}^{c}\sum_{k=1}^{n}|(u_\alpha)_{ik} - (v_\alpha)_{ik}|. \tag{3.49}$$

Nous la reformulons de sorte de montrer qu'il s'agit d'une mesure agrégeant des mesures de compatibilité (stricte) (Définition 13, page 32) construites selon un modèle contraste, avec $\theta = 0$, $\alpha = \beta = 1$ :

$$D_\alpha(U,V) = -\frac{1}{nc}\sum_{i=1}^{c} M_C(U_\alpha^i, V_\alpha^i). \tag{3.50}$$

La distance, symétrique et réflexive, entre les partitions floues $U$ et $V$, est définie par :

$$D_B(U,V) = \int_0^1 D_\alpha(U,V)\, d\alpha \tag{3.51}$$

et calculée comme suit :

$$D_B^\alpha(U,V) = \sum_{i=1}^{|A|} D_{\alpha_i}(U,V)\ (\alpha_i - \alpha_{i-1}) \tag{3.52}$$

où $A = \{\alpha_i\}$ est l'ensemble des degrés d'appartenance distincts de $U$ et $V$, tels que $\alpha_i > \alpha_{i-1}$, avec la convention $\alpha_0 = 0$. On notera $I_B^\alpha(U,V) = 1 - D_B^\alpha(U,V)$ l'indice dérivé. L'auteur ne précise pas de méthode d'appariement particulière entre les clusters de $U$ et ceux de $V$, mais il est tout à fait possible d'adapter la méthode Hongroise.

**Exemple 28.** *Considérons la partition floue* $U'_f = \begin{pmatrix} 0.8 & 0.3 & 0.7 & 0.6 & 0.2 \\ 0.2 & 0.7 & 0.3 & 0.4 & 0.8 \end{pmatrix}$ *de même taille que la partition* $U_f$ *(page 42). L'ensemble des valeurs de coupes à réaliser est* $A = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8\} = \{0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9\}$. *On calcule alors pour chaque élément de A les valeurs de* $D_\alpha(U_f, U'_f)$ *par (3.49). La distance* $D_B^\alpha(U_f, U'_f)$ *vaut 0.1, par (3.52), de sorte que l'indice dérivé vaut quant à lui* $ID_B^\alpha(U_f, U'_f) = 0.9$, *indiquant une forte concordance.*

#### 3.3.2.2 Autres mesures

##### 3.3.2.2.a Acciani et al.

Pour la comparaison de deux partitions floues $U$ et $V$, Acciani et al. [2003] proposent un algorithme original basé sur la construction de deux graphes pondérés non-orientés.

Dans un premier temps, $m = |c - r|$ clusters vides sont adjoints à la partition présentant le moins de clusters. Puis pour chaque partition sont construits les graphes complets $G_U(U, W^U)$ et $G_V(V, W^V)$ dont les sommets sont les clusters de $U$ et $V$, et où $W^U$ et $W^V$ sont les matrices de pondération associées aux arcs de terme général $w_{ij}^U$ et $w_{ij}^V$. Pour $W^U$, ces termes sont définis par :

$$w_{ij}^U = \frac{1}{n} \sum_{k=1}^{n} (u_{ik} \times u_{jk}). \tag{3.53}$$

Dans un second temps, un algorithme de classification hiérarchique est utilisé pour apparier les clusters de $U$ et de $V$, selon la distance euclidienne entre leurs centres respectifs. À partir de cette association, on définit la matrice de permutation $P_\sigma$ et la mesure :

$$MC(U, V) = || W^U - P_\sigma W^V P_\sigma^{-1} ||, \tag{3.54}$$

où l'opérateur $|| \cdot ||$ est la norme de Frobenius définie par : $|| U ||^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2$ .

Cette mesure de discordance présente le désavantage de ne pas avoir de borne supérieure clairement définie, limitant son utilisation à celle d'une mesure relative. Par ailleurs, le choix d'utiliser une classification hiérarchique pour réaliser l'appariement entre les clusters plutôt qu'une approche déterministe telle que la méthode Hongroise est étonnante, car en plus de fournir une permutation sous-optimale, elle peut induire des variations entre deux calculs de $MC(U, V)$ pour les mêmes partitions $U$ et $V$.

**Exemple 29.** *Considérons les deux partitions floues $U_f$ et $V_f$ (page 42). Puisque $U_f$ présente moins de clusters que $V_f$, on lui adjoint $m = c - r = 1$ cluster vide, de sorte d'obtenir la partition dégénérée* $\widetilde{U_f} = \begin{pmatrix} 0.9 & 0.2 & 0.8 & 0.7 & 0.3 \\ 0.1 & 0.8 & 0.2 & 0.3 & 0.7 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$, *puis on calcule les matrices des poids de $\widetilde{U_f}$ et de $V_f$ :*

$$W^{\widetilde{U_f}} = \begin{pmatrix} 2.070 & 0.830 & 0.000 \\ 0.830 & 1.270 & 0.000 \\ 0.000 & 0.000 & 0.000 \end{pmatrix} \text{ et } W^{V_f} = \begin{pmatrix} 0.820 & 0.440 & 0.340 \\ 0.440 & 0.910 & 0.350 \\ 0.340 & 0.350 & 1.010 \end{pmatrix}.$$

*La permutation appariant au mieux les clusters de $\widetilde{U_f}$ et $V_f$ est alors calculée, et l'on obtient :* $P_{(\sigma)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$.

*Enfin, par (3.54) on calcule $MC(U_f, V_f) = 1.779$. Puisque cette mesure n'est pas bornée, il est difficile de discuter ce résultat isolément si bien que par un calcul analogue, on obtient $MC(U_f, W_f) = 1.9586 > MC(U_f, V_f) = 1.779$, ce qui est cohérent avec l'intuition.*

#### 3.3.2.2.b Di Nuovo et Catania

L'indice DNC, qui tire son nom des initiales de ses auteurs Di Nuovo et Catania [2007], est une mesure relative pour la comparaison d'une partition floue $U$ avec une partition de référence stricte $V$. Il s'appuie sur l'idée d'incertitude utilisée lors d'un clustering : plus faible est la proximité entre le plus grand degré d'appartenance $u_{(1)k}$ d'un élément à un cluster et le deuxième plus grand degré d'appartenance $u_{(2)k}$ à un autre cluster, plus faible est la confiance à accorder lors de l'association stricte de cet élément à un cluster. Ainsi, les auteurs définissent un degré de confiance :

$$\Delta_k = u_{(1)k} - u_{(2)k}. \tag{3.55}$$

On reconnait la mesure définie par Ha [1997] pour la sélection de classes en classification supervisée permettant de rejeter pour ambiguïté.

**Proposition 1.** *Nous proposons d'étendre le degré de confiance tel que défini par (3.55) en prenant pour $\Delta_k$ n'importe quelle mesure d'ambiguité $\Phi(\mathbf{u}_k)$, parmi lesquelles toute celles décrites dans [Le Capitaine et Frélicot, 2012].*

À partir de ce degré de confiance, les auteurs définissent un degré de précision :

$$A_k = \begin{cases} \Delta_k & \text{si } u_{ik}^h = v_{ik}\ \forall\ i \in \{1, ..., c\} \\ -\Delta_k & \text{sinon} \end{cases}$$

où $u_{ik}^h$ est le terme général de $U_h$, la partition stricte obtenue par max-défuzzification de $U$. Cette définition implique un appariement des clusters de $U_h$ avec les classes de $V$, pour lequel les auteurs ne proposent aucune méthode. Encore une fois, on pourra utiliser la méthode Hongroise.

Trois ensembles flous sont ensuite introduits :

$$good(A_k) = \begin{cases} 1 & \text{si } A_k \geq \alpha \\ \frac{A_k^2}{\alpha} & \text{si } 0 < A_k < \alpha \\ 0 & \text{si } A_k \leq 0 \end{cases}, bad(A_k) = \begin{cases} 1 & \text{si } A_k \leq -\alpha \\ \frac{\sqrt{A_k}}{\alpha} & \text{si } -\alpha < A_k < 0 \\ 0 & \text{si } A_k \geq 0 \end{cases} \text{ et}$$

$$uncertainty(A_k) = \begin{cases} 0 & \text{si } |A_k| \geq \alpha \\ |A_k| & \text{si } A_k < \alpha \end{cases}.$$

Le paramètre $\alpha$ est un seuil définissant un intervalle d'incertitude autour de 0. Les frontières des ensembles $good$ et $bad$ ont été choisies de sorte de faire preuve d'un certain pessimisme concernant la certitude d'une bonne classification et a contrario, d'un plus grand laxisme sur une probable méprise au sujet d'un mauvais classement.

Finalement, l'indice DNC est défini comme suit :

$$DNC(U, V) = I_a \times (1 - I_u)$$

où $I_a$ et $I_u$ sont deux indices de précision et d'incertitude :

$$I_a = \frac{\sum_k (good(A_k) - bad(A_k))}{n} \quad I_u = \frac{\sum_k uncertainty(A_k)}{n}$$

Cet indice prend sa valeur maximale lorsque $U_h \equiv V$ et $-1 \leq DNC \leq 1$.

**Exemple 30.** *Considérons la partition floue $V_f$ (page 42) et la partition stricte $W_h$ (page 18). À partir de $V_f$, on obtient l'ensemble : $A = \{0.7, -0.5, 0.5, 0.3, 0.5\}$.*
*En choisissant $\alpha = 0.4$, on obtient les ensembles :*

$good = \{1, 0, 1, 0.225, 1\}$, $bad = \{0, 1, 0, 0, 0\}$ *et* $uncertainty = \{0, 0, 0, 0.3, 0\}$.

*À partir d'eux, on calcule les indices de précision et d'incertitude :*

$I_a = \frac{1-1+1+0.225+1}{5} = \frac{2.225}{5} = 0.445$ *et* $I_u = \frac{0.3}{5} = 0.06$

*Finalement l'indice vaut : $DNC(V_f, U_h) = 0.445 \times (1 - 0.06) = 0.418$. Sachant que la valeur forte de $\alpha$ est forte et que DNC prend ses valeurs dans $[-1, 1]$, la partition $V_f$ présente au regard de ce résultat une concordance certaine avec $W_h$.*

### 3.4 A propos de la complexité

Toutes les mesures revues dans ce chapitre ne sont pas égales en terme de temps de calcul et d'espace mémoire nécessaire. Bien évidemment, une telle considération est cruciale lorsque l'on s'intéresse à la comparaison de partitions d'un ensemble de données $X$ de grande taille, et plus encore lorsqu'il s'agit de comparer entre elles plus de deux partitions. La Table 3.14 recense les complexités asymptotiques en temps et en espace affichées par les mesures de comparaison non-strictes présentées dans ce chapitre. Ces complexités ont été calculées sous l'hypothèse que $n \gg max(c, r)$. On remarque que les méthodes orientées individus, qui s'appuient sur une comparaison deux à deux des $n$ individus de l'ensemble $X$ considéré, sont plus coûteuses à la fois en temps de calcul mais aussi en espace, puisque leur calcul nécessite le stockage de $n^2$ valeurs[5]. Les méthodes orientées clusters, qui quant à elles s'attachent à comparer deux à deux les $c \times r$ couples de clusters sont évidemment moins coûteuses en espace. On notera que les méthodes utilisant des normes triangulaires peuvent aussi être accélérées en utilisant leurs fonctions génératrices, comme discuté à la page 48.

### 3.5 Sur la comparaison de partitions issues de différents espaces

Parmi les diverses mesures non-strictes revues dans ce chapitre, nous avons vu que certaines sont définies par leurs auteurs pour la comparaison de deux partitions de $\mathbb{M}_{scn}$ tandis que d'autres se retreignent à la comparaison de deux partitions floues de $\mathbb{M}_{fcn}$ et qu'enfin certaines se limitent à la comparaison d'une partition non-stricte

5. Réductible à $q = \frac{n(n-1)}{2}$ en ne stockant que les valeurs pour les paires d'individus.

TABLE 3.14 – Complexités en temps et en espace des mesures revues en fonction de deux partitions de taille $c \times n$ et $r \times n$.

| Mesure de comparaison | Complexité en temps | Complexité en espace |
|---|---|---|
| Campello (page 53) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Anderson (page 56) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Ceccarelli (page 58) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Borgelt (page 60) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Brouwer (page 63) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| FTD (page 64) | $\mathcal{O}(n)$ | $\mathcal{O}(max(c,r)^2)$ |
| Hüllermeier et al. (page 65) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Beringer (page 68) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Runkler (page 70) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Bodjanova (page 70) | $\mathcal{O}(n)$ | $\mathcal{O}(c^2)$ |
| Acciani (page 71) | $\mathcal{O}(n)$ | $\mathcal{O}(max(c,r)^2)$ |
| DNC (page 73) | $\mathcal{O}(n)$ | $\mathcal{O}(n)$ |

floue et/ou possibiliste avec une partition stricte de référence. La Table 3.15 compile ainsi les différents espaces sur lesquels sont définis chaque mesure. Elle recense aussi pour chaque mesure les espaces sur lesquels une comparaison est techniquement possible, mais n'a pas été prévue ou a été critiquée par son auteur.

Comme nous l'avons discuté au Chapitre 2, en particulier page 8, les différents espaces de partitions non-strictes que sont les espaces des partitions floues et possibilistes portent chacun une sémantique propre si bien qu'en dépit du fait que $\mathbb{M}_{fcn} \subset \mathbb{M}_{pcn}$, nous pensons que la comparaison de deux partitions, l'une floue, l'autre possibiliste, ne doit pas être envisagée. En effet, on ne peut considérer une partition floue comme une partition possibiliste particulière. Dans le cas d'une partition floue, rappelons que les degrés modélisent l'imprécision associée à l'appartenance de l'individu aux clusters, tandis que dans le cas d'une partition possibliste, ces mêmes degrés modélisent la typicalité de ce même individu aux clusters par comparaison à des prototypes. Ainsi, le vecteur d'appartenance possibiliste $\mathbf{u}^p = {}^t(0.5\ 0.5)$ exprime une demi-compatibilité de l'individu considéré à chacun des deux clusters, tandis que le même vecteur d'appartenance flou $\mathbf{v}^f = {}^t(0.5\ 0.5)$ peut quant à lui exprimer une imprécision équi-partagée entre les deux clusters, ou bien une compatibilité maximale avec les deux clusters. Comparer deux partitions $U_p$ et $V_f$ dont ces vecteurs en seraient des colonnes à l'aide des mesures présentées dans ce chapitre n'a pas de sens puisqu'ils coïncideraient. Ceci justifie l'absence d'un tel scénario dans la Table 3.15 et dans les expérimentations menées au Chapitre 4.

TABLE 3.15 – Espaces de définition des mesures revues

| Mesure de comparaison | $\mathbb{M}_{hcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}_{fcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}^{\leq}_{pcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}^{>}_{pcn}$ | $\mathbb{M}_{fcn}$ | $\mathbb{M}^{\leq}_{pcn}$ | $\mathbb{M}^{>}_{pcn}$ |
|---|---|---|---|---|---|---|---|
| Campello (page 53) | $\otimes$ | $\otimes$ | $\otimes$ | $\otimes$ | $\times$ | $\times$ | $\times$ |
| Anderson (page 56) | $\ominus$ | $\ominus$ | $\ominus$ | $\ominus$ | $\ominus$ | $\ominus$ | $\ominus$ |
| Ceccarelli (page 58) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| Borgelt (page 60) | $\otimes$ | $\otimes$ | $\times$ | | $\otimes$ | $\times$ | |
| Brouwer (page 63) | $\otimes$ | $\otimes$ | $\times$ | | $\otimes$ | $\times$ | |
| FTD (page 64) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| Hüllermeier et al. (page 65) | $\otimes$ | $\otimes$ | $\times$ | | $\otimes$ | $\times$ | |
| Beringer (page 68) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| Runkler (page 70) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| Bodjanova (page 70) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| Acciani (page 71) | $\times$ | $\times$ | $\times$ | $\times$ | $\otimes$ | $\times$ | $\times$ |
| DNC (page 73) | $\times$ | $\otimes$ | $\times$ | $\times$ | | | |

$\otimes$ : La mesure est définie par ses auteurs pour une telle comparaison.

$\ominus$ : La mesure est définie par ses auteurs pour une telle comparaison mais présente un défaut de construction qui empêche de l'utiliser l'état.

$\times$ : La mesure peut techniquement être utilisée pour une telle comparaison.

## 3.6 Conclusion

Dans ce chapitre, nous avons dressé un état de l'art espéré exhaustif des mesures de comparaison de partitions non-strictes, selon la même approche taxonomique que celle adoptée au Chapitre 2, fondée sur la distinction entre les approches orientées individus et les approches orientées clusters. Nous avons de plus complété cette taxonomie en distinguant les approches étendant aux cas non-stricts certaines mesures strictes historiques de celles proposant une approche directe, dite native, ne reposant sur aucune construction stricte de la littérature. Enfin, nous avons identifié parmi ces mesures celles généralisant des mesures strictes, c'est-à-dire se ramenant à l'une de ces dernières lorsqu'utilisées pour la comparaisons de partitions strictes.

Par ailleurs, les propriétés métriques de chacune des approches revues ont été discutées et illustrées par l'exemple à l'aide de trois mêmes partitions utilisées tout au long du chapitre. Nous avons ainsi mis en évidence les avantages et les inconvénients de certaines propositions, mais aussi mis au jour certains liens existant entre plusieurs d'entre elles. Une discussion menée à propos des complexités en temps et en espace de chacune de ces dernières montre aussi que les approches orientées clusters se prêtent mieux à la comparaison de partitions caractérisant un grand nombre d'individus. Enfin, nous avons discuté de la comparaison de partitions issues des différents espaces strict, flou et possibiliste, et identifié les scénarios pour lesquels chaque mesure est définie ou utilisable. De ce point de vue, nous n'avons pas observé de prédominance d'une catégorie de mesures (orientées clusters, individus, natives ou non).

Il ressort de cette étude plusieurs points que nous voulons souligner :

- Tout d'abord, nous avons montré que peu de mesures satisfont la propriété de réflexivité, pourtant fortement souhaitable, pour la comparaison de deux partitions strictes. Nous proposons ainsi au chapitre suivant deux nouvelles mesures, l'une orientée individus, l'autre orientée clusters satisfaisant cette propriété.
- Plus généralement, on constate d'ailleurs que la littérature, encore jeune et peu abondante dans le domaine de la comparaison de partitions non strictes, peine à s'accorder sur un ensemble de propriétés théoriques communes qui permettrait pourtant de passer au même crible toute mesure, si bien qu'il est aujourd'hui très difficile de comparer formellement toutes les propositions existantes.
- La définition d'un cadre théorique intégrant tout ou partie des mesures existantes représente donc un enjeu important. En ce sens, la taxonomie adoptée dans ce chapitre met à jour de fortes similitudes entre plusieures propositions orientées clusters, si bien que la définition d'un cadre unifiant ces dernières est

tout à fait envisageable à court terme. De la même manière, les ressemblances constatées entre les diverses approches orientées individus nous ont conduit à dessiner les prémisses d'un formalisme que nous présentons au chapitre suivant.

– Enfin, la comparaison d'une partition floue avec une partition possibliste est un problème mathématiquement ouvert, mais l'est-il sémantiquement ? C'est un obstacle majeur qui nous a fait déconsidérer ce cas de comparaison de partitions non strictes pour nos propositions exposées au chapitre suivant.

# CHAPITRE 4 :

# Contributions

**Résumé :** *Ce chapitre a pour objectif de présenter nos contributions au domaine de la comparaison de partitions non strictes. Y sont décrites trois nouvelles mesures pour chacune desquelles sont menées plusieurs expérimentations spécialement conçues pour en étudier le comportement et le confronter à celui d'autres propositions de la littérature. Enfin, la dernière partie de ce chapitre esquisse les prémisses d'un cadre unifiant les mesures de comparaisons orientées individus.*

## 4.1 Protocoles expérimentaux

Pour chacune des contributions décrites ci-après :

- une extension de mesures strictes orientée individus (page 85),
- une mesure native orientée individus (page 96), et
- une mesure native orientée clusters (page 105),

est présentée une série de résultats obtenus pour différentes expérimentations numériques sur des partitions :

- non strictes synthétiques – **E1**, page 80
- obtenues par clustering de jeux de données
  - synthétiques : partitions strictes – **E2**, page 81, et non strictes – **E3**, page 83
  - réelles[1] : partitions non strictes – **E4**, page 83.

Ces expérimentations ont été conçues pour se compléter, explorant chacune un aspect différent de la problématique de la comparaison de partitions et permettant ainsi de révéler au mieux les comportements des mesures considérées. Ainsi, **E1** et **E3**

1. Extraites de la base de l'UCI Machine Learning Repository [Frank et Asuncion, 2010]

décrivent l'influence des degrés d'appartenance des partitions, tandis que **E2** et **E4** s'attachent à décrire l'influence du nombre de clusters des partitions à comparer sur les mesures considérés. Proposer les mêmes expérimentations à travers tout le présent chapitre présente l'avantage de ne pas surcharger les figures tout en laissant au lecteur la possibilité de comparer les mesures d'une section à l'autre.

Nous nous attachons par ailleurs à étudier le comportement de nos propositions dans le cas de la comparaison de partitions provenant de différents espaces, en évitant toutefois le cas de la comparaison d'une partition floue à une partition possibiliste pour les raisons expliquées au Chapitre 3, page 74. Nous cherchons à comparer nos propositions aux mesures les plus connues de la littérature, selon la philosophie suivante : chaque proposition n'est comparée qu'à des mesures qui lui sont directement comparables. Ainsi par exemple, notre première proposition qui ne satisfait pas la propriété de réflexivité *(I-1)* n'est comparée qu'à des mesures qui elles-mêmes ne satisfont pas cette propriété. Enfin, chaque mesure reposant sur ses propres paramètres, nous nous sommes attachés à en choisir un ensemble restreint assurant une comparaison la plus impartiale possible qui permet quand même au lecteur d'observer leur influence.

### 4.1.1 Partitions non-strictes synthétiques (E1)

Dans cette expérimentation **E1**, on considère deux ensembles $\mathcal{S}$ et $\mathcal{S}'$, tous deux composés de neuf partitions de même taille ($c = 3, n = 4$). L'ensemble $\mathcal{S}$ regroupe ainsi :

- une partition stricte $U_1$,
- huit partitions floues $\{U_2, ..., U_9\}$ max-compatibles avec $U_1$ et telles que $U_{i+1}$ est plus floue que $U_i$, $\forall i \in \{1, ..., 8\}$ selon l'*entropie de partition* (ou *Partition Entropy*) *PE* proposée par Bezdek [1981], qui mesure le degré de flou au sein d'une partition :

$$PE(U) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\ \log(u_{ik}). \tag{4.1}$$

  Ces valeurs sont données à la Table 4.1 pour chacune des partitions de $\mathcal{S}$.

De la même manière, l'ensemble $\mathcal{S}'$ regroupe les partitions suivantes :

- une partition stricte $U'_1 = U_1$,
- huit partitions possibilistes $\{U'_2, ..., U'_9\}$ de $\mathbb{M}^{\leq}_{pcn}$, max-compatibles avec $U'_1$ et telles que construites à partir des précédentes selon $U'_i = \frac{1}{2}U_i$.

TABLE 4.1 – Degré de flou des partitions de l'ensemble $S$

| $U_i$ | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ | $U_7$ | $U_8$ | $U_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $PE(U_i)$ | 0 | 0.3816 | 0.5965 | 0.9144 | 1.1332 | 1.3025 | 1.4101 | 1.4511 | 1.4614 |

Les partitions de chaque ensemble $S$ et $S'$ sont ainsi choisies de sorte d'être triées par ordre décroissant[2] selon leur proximité avec la partition stricte $U_1$.

Pour chaque mesure, les résultats obtenus pour la comparaison des $9 \times 9 = 81$ couples de partitions $(U_i, U_j)$ sont présentés sous la forme d'une mosaïque de valeurs, où les partitions sont ordonnées en ligne et en colonne selon leur indice, comme illustré à la Figure 4.1. Plus une case de la mosaïque est claire, plus la valeur de la mesure de comparaison est grande pour le couple $(U_i, U_j)$ correspondant, et plus les deux partitions concordent. Comme toutes les mesures de comparaison étudiés dans ce chapitre sont symétriques (propriété (*I-3*)), seule la partie supérieure de chaque mosaïque est présentée.

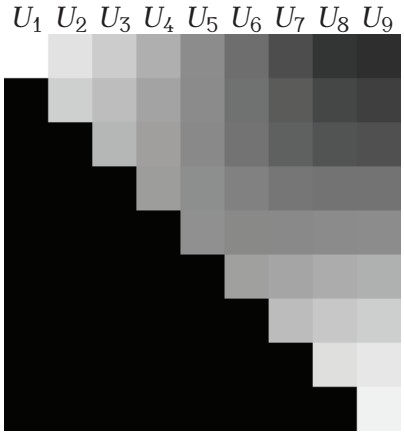



FIGURE 4.1 – **E1.** Une mosaïque de valeurs de comparaison pour les partitions de $S$.

### 4.1.2 Partitions obtenues par clustering

#### 4.1.2.1 Jeux de données synthétiques

##### 4.1.2.1.a Partitions strictes (E2)

L'objectif de cette expérimentation **E2**, est de comparer deux partitions strictes d'un même jeu de données synthétique bidimensionnel composé de $k^\star = 5$ clusters gaussiens isotropes, centrés en $(-1, 6)$, $(1, 1)$, $(-1, 2)$, $(5, 4)$ et $(3, 5)$, partageant le même

2. L'entropie des partitions de $S'$ ne sont pas données. Elles ne sont pas proportionnelles à celles de $S$ mais la relation d'ordre demeure.

écart-type $\Sigma = \frac{1}{2}Id$. Deux paires de ces clusters se touchent légèrement et sont bien séparées du cinquième cluster, comme le montre la première sous-figure de la Figure 4.2.

Les autres sous-figures présentent les partitions strictes $V_k$ obtenues avec l'algorithme des $k-Moyennes$ (implémentation Matlab), pour $k = 2, 3, ..., 12$. Chaque partition $V_k$ est comparée pour chaque indice considéré avec la partition $U_{k^\star}$ de référence, et les valeurs $I(U_{k^\star}, V_k)$ sont tracées pour chaque indice en fonction de $k$.

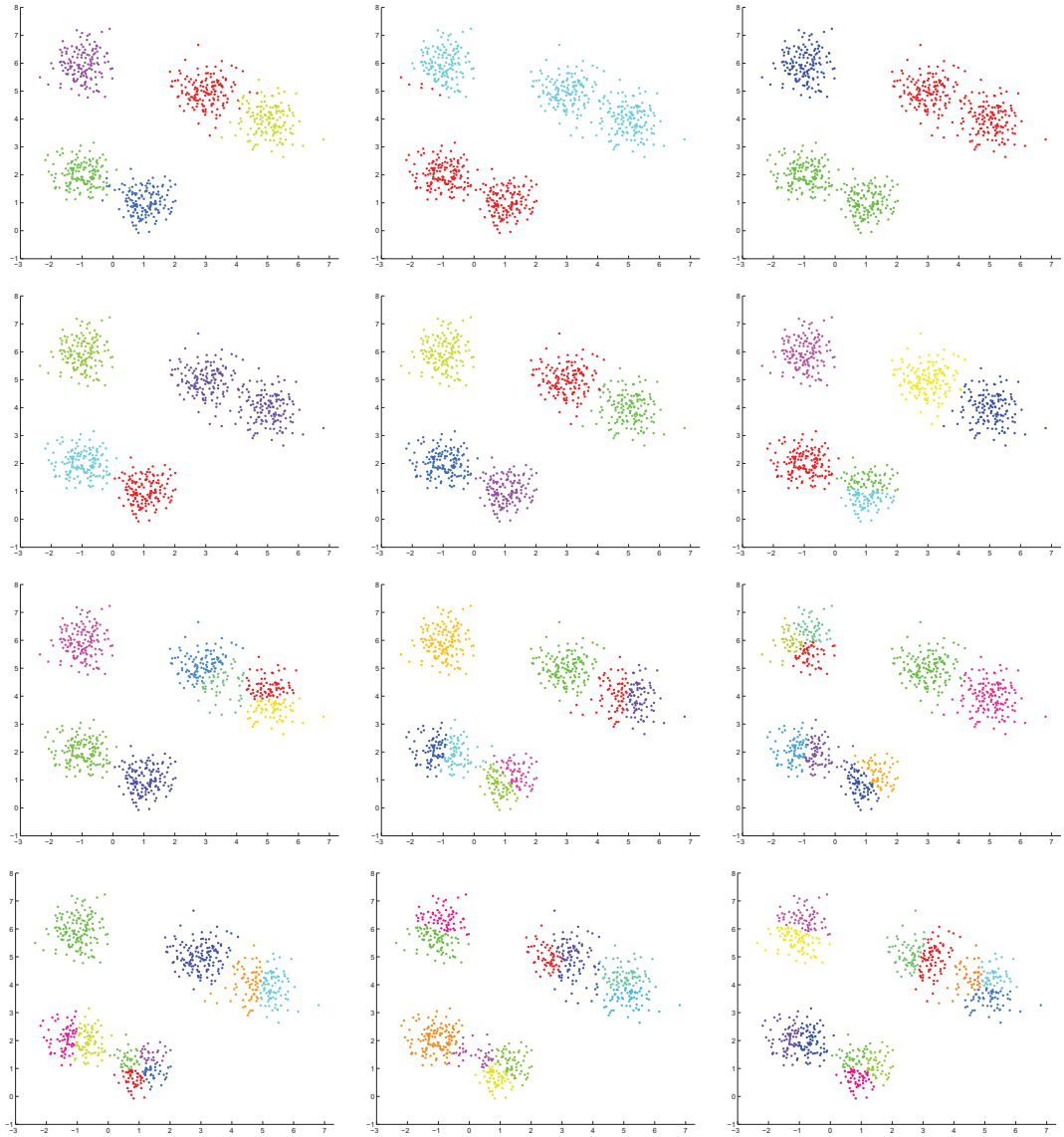


FIGURE 4.2 – **E2.** De haut en bas et de gauche à droite : partition véritié-terrain $U_{k^\star}$ et partitions strictes $V_k$ $(k = 2, ..., 12)$ obtenues avec l'algorithme des $k-Moyennes$

#### 4.1.2.1.b Partitions non-strictes (E3)

Inspirée des travaux de Ceccarelli et Maratea [2008], cette expérimentation **E3** s'attache à comparer une partition stricte de référence $U_{k^\star}$ à deux collections de partitions non-strictes, les unes floues et les autres possibilistes. Dix jeux de données tri-dimensionnel ont été générés pour des écarts-types $\sigma = \{\frac{1}{j} : j = 10, 9, ..., 1\}$, de telle sorte que le chevauchement entre les clusters de chaque jeu de données croisse, comme le montre la Figure 4.3.

L'implémentation *Matlab* des algorithmes FCM (*C-Moyennes Floues*, Annexe A.1, page 145) et PCM (*C-Moyennes Possibilistes*, voir Annexe A.2, page 147) ont été lancées[3] pour chacun de ces jeux de données afin de produire respectivement dix 3-partitions floues $V_\sigma \in \mathbb{M}_{fcn}$ et dix 3-partitions possibilistes $V'_\sigma \in \mathbb{M}^{\leq}_{pcn}$. Pour chaque indice, les partitions de chacun de ces ensembles sont comparées une à une avec la partition stricte $U_{k^\star}$ de référence. En résultent deux courbes donnant la valeur de l'indice considéré $I(U_{k^\star}, V_\sigma)$ et $I(U_{k^\star}, V'_\sigma)$ en fonction de $\sigma$ pour les ensembles de partitions floues et possibilistes.

### 4.1.2.2 Partitions non-strictes de jeux de données réelles (E4)

Cette expérimentation **E4** est conduite sur plusieurs jeux de données issus de la base de l'UCI (*UCI Machine Learning Repository* [Frank et Asuncion, 2010]) présentant des caractéristiques variées en termes du nombre $n$ d'individus, du nombre $p$ d'attributs, du nombre de classes $c^\star$ connu et du degré de chevauchement entre les classes, estimé grossièrement via une ACP (*Analyse en Composantes Principales*, [Jolliffe, 2002]).

Pour chaque jeu de données, les algorithmes FCM et PCM ont été utilisés (même implémentation, mêmes paramètres d'exécution que dans l'expérimentation **E3**, page 83) pour produire respectivement une $c^\star$-partition de référence floue $R_{c^\star} \in \mathbb{M}_{fcn}$ et une $c^\star$-partition de référence possibiliste $R'_{c^\star} \in \mathbb{M}^{\leq}_{pcn}$, ainsi que deux collections composées respectivement de onze $c$-partitions floues $U_c$ et onze $c$-partitions possibilistes $U'_c \in \mathbb{M}^{\leq}_{pcn}$, pour $c$ variant de 2 à 12. Comme pour l'expérimentation précédente, chaque partition $U_c$ (respectivement $U'_c$) est comparée à la partition de référence $R_{c^\star}$ (respectivement $R'_{c^\star}$) pour produire pour chaque indice une courbe des valeurs $I(R_{c^\star}, U_c)$ (respectivement $I(R'_{c^\star}, U'_c)$) en fonction de $c$.

3. L'exposant de flou, le seuil de convergence et le nombre maximum d'itérations sont définis pour chaque algorithme à 2, $10^{-3}$ et 1000, respectivement.

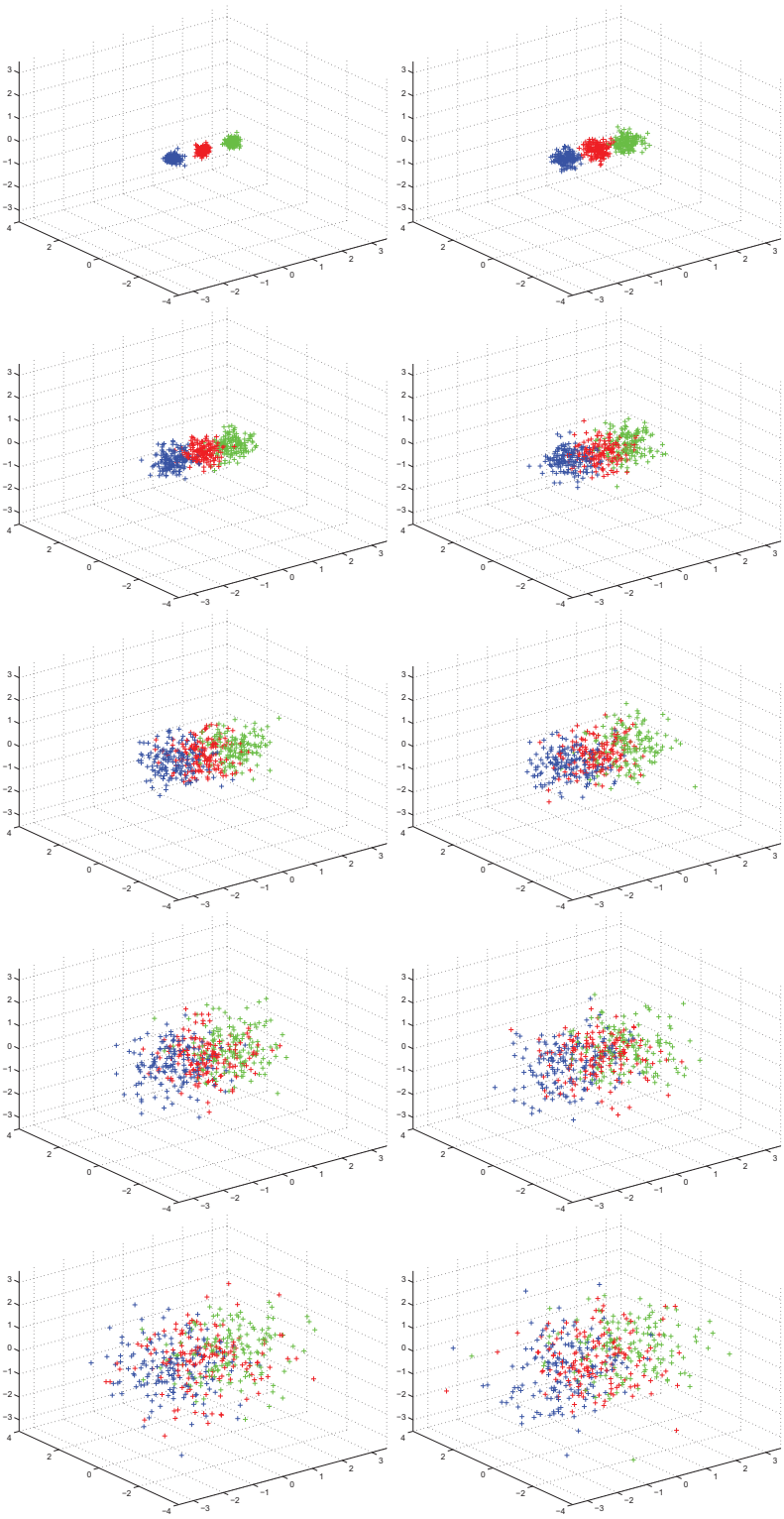


FIGURE 4.3 – **E3.** Dix partitions vérité-terrain $U_{k^\star}$ : $\sigma = \{\frac{1}{j} : j = 10, 9, ..., 1\}$

Les résultats obtenus sur l'ensemble des jeux considérés s'étant révélés convergents en termes de comportements observés pour chaque indice étudié, nous nous restreindrons, par souci d'espace et d'équilibre de ce mémoire, à ne présenter que les résultats obtenus pour les quatre jeux de données suivants :

- Fisher iris ($n = 150, p = 4, c^\star = 3$, faible chevauchement entre deux classes),
- Pima diabetes ($n = 768, p = 8, c^\star = 2$, fort chevauchement entre les deux classes),
- Italian wine ($n = 178, p = 13, c^\star = 3$, faible chevauchement entre les trois classes),
- Pageblocks ($n = 5473, p = 10, c^\star = 5$, fort chevauchement entre plusieurs classes).

## 4.2 Une extension de mesures strictes orientée individus

Dans cette première contribution, nous proposons de généraliser à toute t-norme la normalisation des matrices de coïncidence floues réalisée par Brouwer [2009a] pour la t-norme produit $\top_P$ et décrite page 63. Rappelons que cette normalisation permet de ramener à 1, pour toute partition floue $U$ de $\mathbb{M}_{fcn}$, les termes diagonaux $\psi_{U,kk}$ de sa matrice de coïncidence floue (3.25) par la transformation suivante (3.27) :

$$\phi^{\top_P}_{U,kl} = \frac{\psi^{\top_P}_{U,kl}}{\sqrt{\psi^{\top_P}_{U,kk}}\sqrt{\psi^{\top_P}_{U,ll}}},$$

définissant ainsi une nouvelle matrice de coïncidence normalisée $\Phi_U$.

### 4.2.1 Proposition

Afin d'obtenir, à partir une matrice de coïncidence floue $\Psi^{\top}_U$, une matrice $\Phi^{\top}_U$, normalisée dans le sens où tous ses termes diagonaux $\phi^{\top}_{U,kk}$ sont égaux à 1 pour n'importe quelle t-norme, il suffit de trouver une fonction $K_\top(a) : [0,1] \to [0,1]$ telle que

$$\begin{aligned} \frac{a}{\top\left(K_\top(a), K_\top(a)\right)} &= 1 \\ \Leftrightarrow \top\left(K_\top(a), K_\top(a)\right) &= a. \end{aligned} \tag{4.2}$$

Étant donné un couple $(\top, K_\top)$, (3.27) est ainsi facilement généralisé par :

$$\phi^{\top}_{U,kl} = \frac{\psi^{\top}_{U,kl}}{\top\left(K_\top\left(\psi^{\top}_{U,kk}\right), K_\top\left(\psi^{\top}_{U,ll}\right)\right)}. \tag{4.3}$$

Cette transformation n'affecte pas les indices si les partitions sont strictes. En effet, si $U \in \mathbb{M}_{hcn}$, alors $\psi^{\top}_{U,kk} = \psi^{\top}_{U,ll} = 1$ et, par conséquent, le dénominateur de (4.3) est

égal à 1 par (4.2). Pour les t-normes basiques du Tableau 3.1, les fonctions $K_\top$ sont faciles à trouver, aussi nous laissons le soin au lecteur de chercher les preuves.

**Théorème 5.** *[Quéré et al., 2010]*
*Pour les t-normes basiques* standard, produit *et* Łukasiewicz, *les fonctions normalisantes $K_\top$ sont respectivement :*

$$K_{\top_M}(a) = a, \tag{4.4}$$
$$K_{\top_P}(a) = \sqrt{a}, \tag{4.5}$$
$$K_{\top_Ł}(a) = \frac{a+1}{2}. \tag{4.6}$$

Nous proposons d'utiliser le générateur additif ou le générateur multiplicatif des t-normes paramétriques pour dériver les fonctions normalisantes $K_\top$.

**Théorème 6.** *[Quéré et al., 2010]*
*Soit une t-norme archimédienne*[4] $\top$, *de générateur additif $f_\top$ ou multiplicatif $g_\top$, la fonction normalisante $K_\top$ telle que $\top(K_\top(a), K_\top(a)) = a$ est*

$$K_\top(a) = \begin{cases} f_\top^{-1}\left(\frac{f_\top(a)}{2}\right) \\ g_\top^{-1}\left(\sqrt{g_\top(a)}\right) \end{cases} \tag{4.7}$$

*où $f_\top^{-1}$ et $g_\top^{-1}$ sont respectivement les pseudo-inverses de $f_\top$ et $g_\top$.*

*Démonstration.* Selon la définition de $\top$, nous avons

$$\top(a,b) = \begin{cases} f_\top^{-1}\left(f_\top(a) + f_\top(b)\right) \\ g_\top^{-1}\left(g_\top(a)\, g_\top(b)\right) \end{cases} \tag{4.8}$$

Ainsi,

$$\begin{aligned} (4.2) &\Leftrightarrow f_\top^{-1}\left(2 f_\top(K_\top(a))\right) = a \\ &\Leftrightarrow f_\top(K_\top(a)) = \frac{f_\top(a)}{2} \\ &\Leftrightarrow K_\top(a) = f_\top^{-1}\left(\frac{f_\top(a)}{2}\right) \end{aligned}$$

ou

$$\begin{aligned} (4.2) &\Leftrightarrow g_\top^{-1}\left(g_\top(K_\top(a))^2\right) = a \\ &\Leftrightarrow K_\top(a) = g_\top^{-1}\left(\sqrt{g_\top(a)}\right) \end{aligned}$$

□

4. Définition 18, page 42.

Les fonctions normalisantes des familles de t-normes paramétriques de la Table 3.2 (page 46), obtenues avec leur générateur additif $f$ sont donnés à la Table 4.2. On constate que dans le cas où les valeurs du paramètre $\lambda$ ramènent la t-norme à l'une des t-normes basiques, et à condition que cette dernière soit archimédienne ($\top_P$, $\top_Ł$, $\top_D$), on en retrouve la fonction normalisante correspondante. Par exemple :

- $K_{\top_{AA_1}}(a) = a^{\frac{1}{2}} = K_{\top_P}(a)$,
- $K_{\top_{H_1}}(a) = \sqrt{a} = K_{\top_P}(a)$,
- $K_{\top_{SS_1}}(a) = \frac{a+1}{2} = K_{\top_Ł}(a)$,
- $K_{\top_{Y_1}}(a) = 1 - \frac{1-a}{2} = K_{\top_Ł}(a)$.

Cette même propriété tient pour les relations entre les familles de t-normes paramétriques elles-mêmes.

Table 4.2 – Fonctions normalisantes des principales familles de t-normes paramétriques de la Table 3.2

| $\top$-norme | $K_\top(a)$ |
|---|---|
| Aczel-Alsina | $a^{\left(\frac{1}{2}\right)^{1/\lambda}}$ |
| Dombi | $\left(1 + \left(\frac{1}{2}\right)^{1/\lambda} \frac{1-a}{a}\right)^{-1}$ |
| Frank | $\frac{\ln\left((\lambda-1)\sqrt{\frac{\lambda^a-1}{\lambda-1}}+1\right)}{\ln\lambda}$ |
| Hamacher | $\frac{\lambda\sqrt{a}}{\sqrt{\lambda+(1-\lambda)a}+(\lambda-1)\sqrt{a}}$ |
| Schweizer-Sklar | $\left(\frac{a^\lambda+1}{2}\right)^{1/\lambda}$ |
| Sugeno-Weber | $\frac{\sqrt{(1+\lambda)(1+\lambda a)}-1}{\lambda}$ |
| Yager | $1 - \frac{1-a}{2^{1/\lambda}}$ |

**Exemple 31.** *Les t-normes de Hamacher et de Dombi sont égales si leur paramètres respectifs sont $\lambda = 0$ et $\lambda = 1$. Du générateur et du pseudo-inverse de $\top_{H_0}$, définis respectivement par $f(a) = \frac{1-a}{a}$ et $f^{-1}(a) = \frac{1}{1+a}$, nous obtenons par (4.7) :*

- $K_{\top_{H_0}}(a) = \frac{1}{1+\frac{1-a}{2a}}$, *qui équivaut à* $K_{\top_{D_1}}(a)$.

Par contraposition, lorsque les fonctions de la Table 4.2 sont indéfinies pour certaines valeurs de $\lambda$ on utilisera les fonctions normalisantes des t-normes auxquelles elles se ramènent le cas échéant.

**Exemple 32.** *La fonction normalisante associée à la famille des t-normes de Sugeno-Weber est indéfinie pour $\lambda = 0$. Toutefois dans ce cas, $\top_{SW_0} = \top_L$, si bien que $K_{\top_{SW_0}}(a) = 1 - \frac{1-a}{2}$, est en fait égal à $K_{\top_L}(a)$.*

Une autre option dans le cas d'indéfinition d'une t-norme paramétrique consiste à passer par le calcul de la limite de la fonction normalisante en la valeur de $\lambda$ problématique.

**Exemple 33.** *Considérons la t-norme de Frank, qui est égale aux t-normes basiques standard $\top_M$, produit $\top_P$ et de Łukasiewicz lorsque $\lambda$ est respectivement proche de 0, 1 et $+\infty$. En utilisant au besoin les séries de Taylor, et en posant $N = (\lambda - 1)\sqrt{\frac{\lambda^a - 1}{\lambda - 1}} + 1$, on peut montrer que :*

- *lorsque $\lambda \to 0$,*
  $N \to \frac{1}{2}(\lambda^a + \lambda) + (\lambda^a - \lambda)\,\varepsilon(\lambda)$,
  $\ln(N) \to a \ln(\lambda)$ *et nous avons donc, comme prévu* $K_{\top_{F_0}}(a) \to a = K_{\top_M}(a)$,
- *lorsque $\lambda \to 1$,*
  $N \to u\sqrt{a + \varepsilon(u)} + 1$ *où* $\lambda = 1 + u$,
  $\ln(N) \to u\sqrt{a}$ *et nous avons donc, comme prévu* $K_{\top_{F_1}}(a) \to \sqrt{a} = K_{\top_P}(a)$,
- *lorsque $\lambda \to +\infty$,*
  $\ln(N) \to \ln(\lambda^{\frac{a+1}{2}})$ *et nous avons donc, comme prévu* $K_{\top_{F_{+\infty}}}(a) \to \frac{a+1}{2} = K_{\top_L}(a)$.

Cette normalisation des matrices de coïncidence floues généralisée à toute t-norme permet de transformer les cardinalités floues $n^{\top}_{\alpha\beta}(\Psi^{\top}_U, \Psi^{\top}_V)$ données par (3.26) par $n^{\top}_{\alpha\beta}(\Phi^{\top}_U, \Phi^{\top}_V)$, calculées à partir des matrices de coïncidences normalisées définies par (4.3). Pour chaque couple $(\top, K_{\top})$, nous pouvons ainsi dériver de nouvelles extensions de toutes les mesures d'appariement existantes, et en particulier les indices de Rand et de Jaccard notés respectivement $RI^{\top}_Q$ et $JI^{\top}_Q$. Notons enfin que pour la t-norme produit $\top_P$, l'invariance en échelle des matrices de coïncidence normalisées (3.28) induit une invariance en échelle de nos indices, si bien que : $\forall \alpha \in \mathbb{R}_+$,

$$RI^{\top_P}_Q(U, \alpha V) = RI^{\top_P}_Q(U, V) \quad \text{et} \quad JI^{\top_P}_Q(U, \alpha V) = JI^{\top_P}_Q(U, V). \tag{4.9}$$

**Exemple 34.** *Considérons la partition-exemple $U_f$ du chapitre 3 (page 42), pour laquelle la matrice de coïncidence normalisée $\Phi^{\top_P}_{U_f}$ selon (3.27) a été donnée à l'Exemple 23. Pour la t-norme paramétrique de Hamacher $\top_{H_0}$, on obtient par (4.3) :*

$$\Psi^{\top_{H_0}}_{U_f} = \begin{pmatrix} 0.781 & 0.183 & 0.672 & 0.571 & 0.245 \\ 0.183 & 0.563 & 0.195 & 0.225 & 0.470 \\ 0.672 & 0.195 & 0.563 & 0.470 & 0.225 \\ 0.571 & 0.225 & 0.470 & 0.391 & 0.228 \\ 0.245 & 0.470 & 0.225 & 0.228 & 0.391 \end{pmatrix}$$

*et*

$$\Phi_{U_f}^{\top_{H_0}} = \begin{pmatrix} 1.000 & 0.272 & 0.997 & 0.982 & 0.420 \\ 0.272 & 1.000 & 0.347 & 0.475 & 0.992 \\ 0.997 & 0.347 & 1.000 & 0.992 & 0.475 \\ 0.982 & 0.475 & 0.992 & 1.000 & 0.584 \\ 0.420 & 0.992 & 0.475 & 0.584 & 1.000 \end{pmatrix}$$

*Les matrices de contingence-paires non normalisée et normalisée, croisant la partition $U_f$ avec elle-même, dérivées valent alors :*

$$M_B^{\top_{H_0}}(U_f, U_f) = \begin{pmatrix} 0.795 & 1.096 \\ 0.1096 & 3.519 \end{pmatrix} \text{ et } M_Q^{\top_{H_0}}(U_f, U_f) = \begin{pmatrix} 4.486 & 0.760 \\ 0.760 & 1.257 \end{pmatrix}.$$

*On constate que la normalisation induit une forte augmentation du terme d'accord $m_{11}$, au détriment des termes de désaccord $m_{10}$ et $m_{01}$, comme souhaité.*

### 4.2.2 Expérimentations

Dans cette partie, nous allons étudier le comportement des indices de Rand $RI_Q^\top$ et de Jaccard $JI_Q^\top$ que nous avons proposés à partir des résultats obtenus pour les expérimentations décrites en début de ce chapitre, page 79. Notre proposition permettant de généraliser tous les indices stricts orientés individus fondés sur les termes de la matrice de contingence-paires, ils s'y ramènent dans le cas de la comparaison de partitions strictes. C'est pourquoi l'expérimentation **E2** portant sur la comparaison de deux partitions strictes obtenues par clustering de jeux de données synthétiques n'est pas présentée ici. Les valeurs de $RI$ et $JI$ (donc de $RI_Q^\top$ et $JI_Q^\top$) qui lui sont relatives sont présentées à la Figure 4.21, page 117.

Notre contribution est confrontée à trois propositions de la littérature dont les propriétés lui sont analogues : les approches orientées individus de Campello [2007] et de Anderson et al. [2010] (dans sa version corrigée par nos soins) décrites respectivement pages 53 et 56, et celle de Borgelt [2006], décrite page 60. Pour toutes ces approches, sont dérivés les indices de Rand et de Jaccard, notés respectivement $RI_{\widetilde{A}_\star}^\top$ et $JI_{\widetilde{A}_\star}^\top$ pour Anderson et al., $RI_C^\top$ et $JI_C^\top$ pour Campello et $RI_B^\top$ et $JI_B^\top$ pour Borgelt. Dans un souci de concision, seuls les résultats obtenus avec deux t-normes (et leur conorme duale le cas échéant) sont présentés ci-après. Nous avons ainsi retenus les norme triangulaire produit $\top_P$ et la norme de Hamacher $\top_{H_{100}}$ pour leur popularité et leur comportement très différent. Les résultats obtenus avec d'autres paramétrages sont donnés en Annexe B. Notons que les indices $RI_{\widetilde{A}_\star}^\top$ et $JI_{\widetilde{A}_\star}^\top$ ne sont calculés qu'avec $\top_P$ puisqu'il n'existe pas aujourd'hui de méthode de calcul pour son approche généralisée à toute t-norme.

#### 4.2.2.1 Partitions non-strictes synthétiques (E1)

Nous présentons et discutons ici les résultats de l'expérimentation **E1** décrite page 80. La Figure 4.4 montre les mosaïques de valeurs de comparaison calculées avec les deux indices $RI^{\top_P}_{\widetilde{A_\star}}$ et $JI^{\top_P}_{\widetilde{A_\star}}$ pour les deux jeux de partitions floues $\mathcal{S}$ et possibilistes $\mathcal{S}'$, tandis que la Figure 4.5 présente les mêmes résultats obtenus avec les autres indices considérés pour d'autres t-normes puisque ceux-ci le permettent.

Tout d'abord, on remarque que tous les indices considérés suivent bien la relation d'ordre établie ad-hoc entre les partitions, chacune des mosaïques présentant une luminosité plus forte dans les zones où les partitions sont proches et très peu floues (coin supérieur gauche) que dans les zones où les partitions sont plus éloignées (coin supérieur droit). Par ailleurs, on remarque que pour tous les indices, les diagonales des mosaïques ne sont pas blanches, indiquant que ces propositions ne préservent pas la propriété de réflexivité *(II-1)*. Ensuite, le faible contraste des mosaïques des indices $RI^{\top}_{\widetilde{A_\star}}$, $RI^{\top}_{C}$ et $RI^{\top}_{B}$ témoigne de leur faible dynamique (propriété *(II-4)*), contrairement à celle de $RI^{\top}_{Q}$. Ainsi, les résultats obtenus pour notre indice $RI^{\top}_{Q}$ montrent que la normalisation augmente de manière significative la dynamique de l'indice $RI^{\top}_{B}$ qu'elle propose de corriger. Les indices de Jaccard tels que nous les proposons présentent quant à eux une meilleure granularité que ceux dérivés des autres approches. En effet, bien que le contraste des mosaïques correspondant à l'indice de Jaccard soit globalement homogène pour toutes les approches, notre proposition fait apparaitre une transition plus douce que les autres de la partition stricte (à gauche de chaque mosaïque) vers la partition la plus floue (à droite). Dans la mesure où les partitions considérées sont toutes max-compatibles entre elles, ce comportement peut-être un avantage pour une utilisation en tant que mesure relative. On remarque aussi que l'influence de la t-norme est moindre pour nos indices, même si elle est bien présente. En effet, l'utilisation de la t-norme $\top_{H_{100}}$, dont le comportement est plutôt drastique

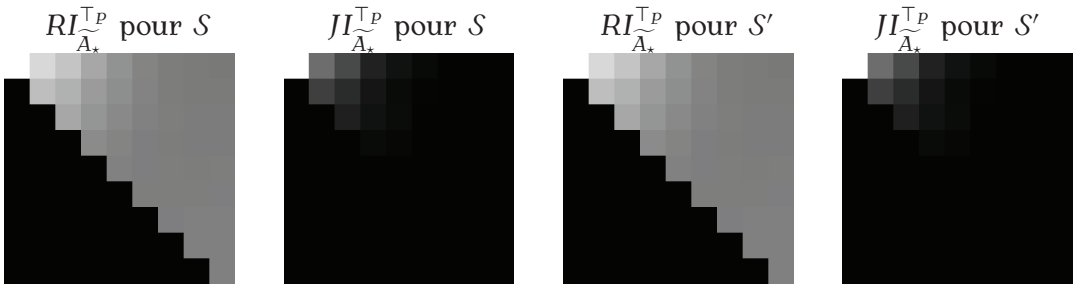



Figure 4.4 – Mosaïques des valeurs de comparaison obtenues avec les extensions de $RI$ et de $JI$ de Anderson et al. pour les partitions floues de $\mathcal{S}$ et possibilistes de $\mathcal{S}'$

FIGURE 4.5 – Mosaïques des valeurs de comparaison obtenues avec les extensions de *RI* et *JI* proposées par Campello, Borgelt et Quéré et al., pour différentes t-normes. Première et deuxième ligne : partitions floues de $\mathcal{S}$. Troisième et quatrième ligne : partitions possibilistes de $\mathcal{S}'$.

(voir Figure 3.3, page 47), rend presque homogène les mosaïques des indices de Borgelt, car elle favorise les fortes valeurs des matrices de coïncidence au détriment des valeurs plus faibles, ce qui a pour incidence de réduire le terme $m_{11}$ de la matrice de contingence-paires au profit du terme $m_{00}$. L'indice $RI_B^{\top_{H_{100}}}$, pour lequel $m_{00}$ contribue à l'accord, ne présente donc que de très fortes valeurs, tandis qu'a contrario, $JI_B^{\top_{H_{100}}}$, pour lequel seul $m_{11}$ mesure l'accord des partitions comparées, ne présente que des valeurs faibles. Notre proposition corrige ce comportement, tout en préservant le comportement induit par $\top_{H_{100}}$, puisque les valeurs des indices obtenues pour les partitions présentant les degrés d'appartenance les plus forts sont légèrement supérieures à celles obtenues avec la t-norme produit $\top_P$. Enfin, il apparaît que les résultats obtenus pour les partitions possiblistes de l'ensemble $\mathcal{S}'$ diffèrent peu de ceux obtenus pour les partitions floues de $\mathcal{S}$, et les mosaïques obtenues avec $RI_Q^{\top_P}$ et $JI_Q^{\top_P}$ pour les partitions possibilistes sont même strictement égales à celles obtenues pour les partitions floues. Ceci résulte de la propriété d'invariance en échelle de nos indices pour $\top_P$ (4.9), puisque les partitions possibilistes de $\mathcal{S}'$ sont liées par construction par un simple facteur d'échelle à celles de $\mathcal{S}$. Pour tous les autres indices, les valeurs obtenues avec les partitions possibilistes sont très légèrement plus faibles, mais les comportements constatés demeurent. C'est pourquoi nous relègueront les résultats obtenus pour les partitions possibilistes des autres expérimentations à l'Annexe B.

#### 4.2.2.2 Partitions obtenues par clustering

##### 4.2.2.2.a Données synthétiques - Partitions non strictes (E3)

Nous présentons et discutons ici les résultats de l'expérimentation **E3**, décrite page 83, dédiée à la comparaison d'une partition stricte de référence avec une collection de partitions non strictes. La Figure 4.6 présente les résultats obtenus pour les partitions floues fournies par l'algorithme FCM. Ceux obtenus pour les partitions possibilistes fournies par l'algorithme PCM sont analogues, si bien qu'ils sont donnés en Annexe B.

On observe tout d'abord que tous les indices considérés prennent leur valeur maximum, très proche de 1, pour $\sigma = 0.1$, ce qui n'est pas étonnant puisque les données sont dans ce cas bien séparées. Les valeurs décroissent ensuite avec $\sigma$, comme attendu, à mesure que le chevauchement entre les clusters se fait plus important. Ensuite, toutes les remarques faites à propos des résultats de l'expérimentation **E1** peuvent ici aussi être formulées. Nos indices de Rand $RI_Q^{\top}$ font ainsi généralement montre d'une meilleure dynamique que les autres. Pour l'indice de Jaccard en revanche, notre proposition se positionne très légèrement derrière vis-à-vis de ce critère, notamment face à $JI_C^{\top_{H_{100}}}$, bien qu'elle conserve un comportement tout à fait correct et cohérent.

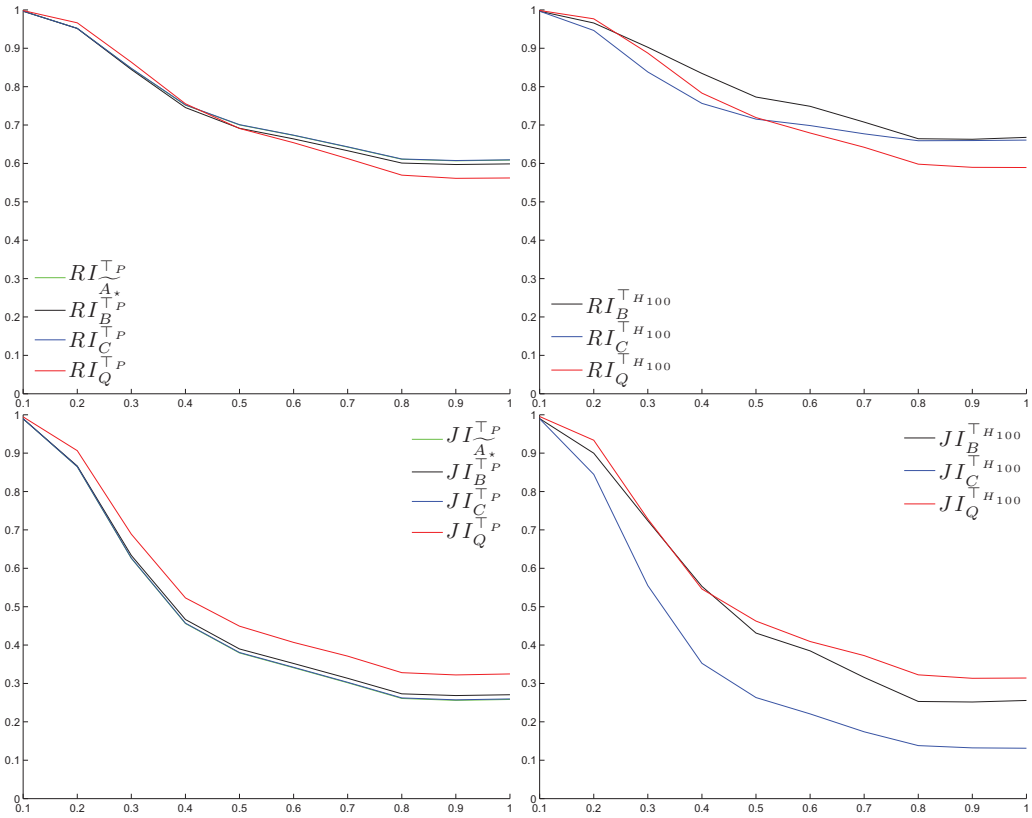



FIGURE 4.6 – Comparaison d'une collection de partitions floues $V_\sigma$ à une partition stricte de référence $U_{k^\star}$ avec les indices de Rand et de Jaccard dérivés selon Anderson et al., Campello, Borgelt et notre proposition.

Ceci s'explique, notamment lorsque l'on considère les résultats obtenus avec $\top_{H_{100}}$, par le fait que la normalisation amoindrit le comportement drastique des t-normes, comme expliqué précédemment. On notera enfin que les indices de Anderson et al. et de Borgelt présentent des comportements similaires. Ceci découle du lien qui les rapproche, que nous avons prouvé (Proposition 2, page 61).

#### 4.2.2.2.b Partitions non-strictes de données réelles (E4)

Nous donnons ici les résultats de l'expérimentation **E4**, décrite page 83, consacrée à la comparaison de partitions non-strictes obtenues par clustering de jeux de données réelles. Pour les partitions floues $U_c$, les résultats sont donnés aux Figures 4.7 (indices de Rand) et 4.8 (indices de Jaccard). Encore une fois, ceux obtenus pour les partitions possibilistes $U'_c$ étant sensiblement analogues sont reportés en Annexe B.

Sur tous les jeux de données, tous les indices prennent leur valeur maximale pour $c = c^\star$, sans toutefois toujours atteindre la valeur 1, même si les partitions $U_{c^\star}$ fournies

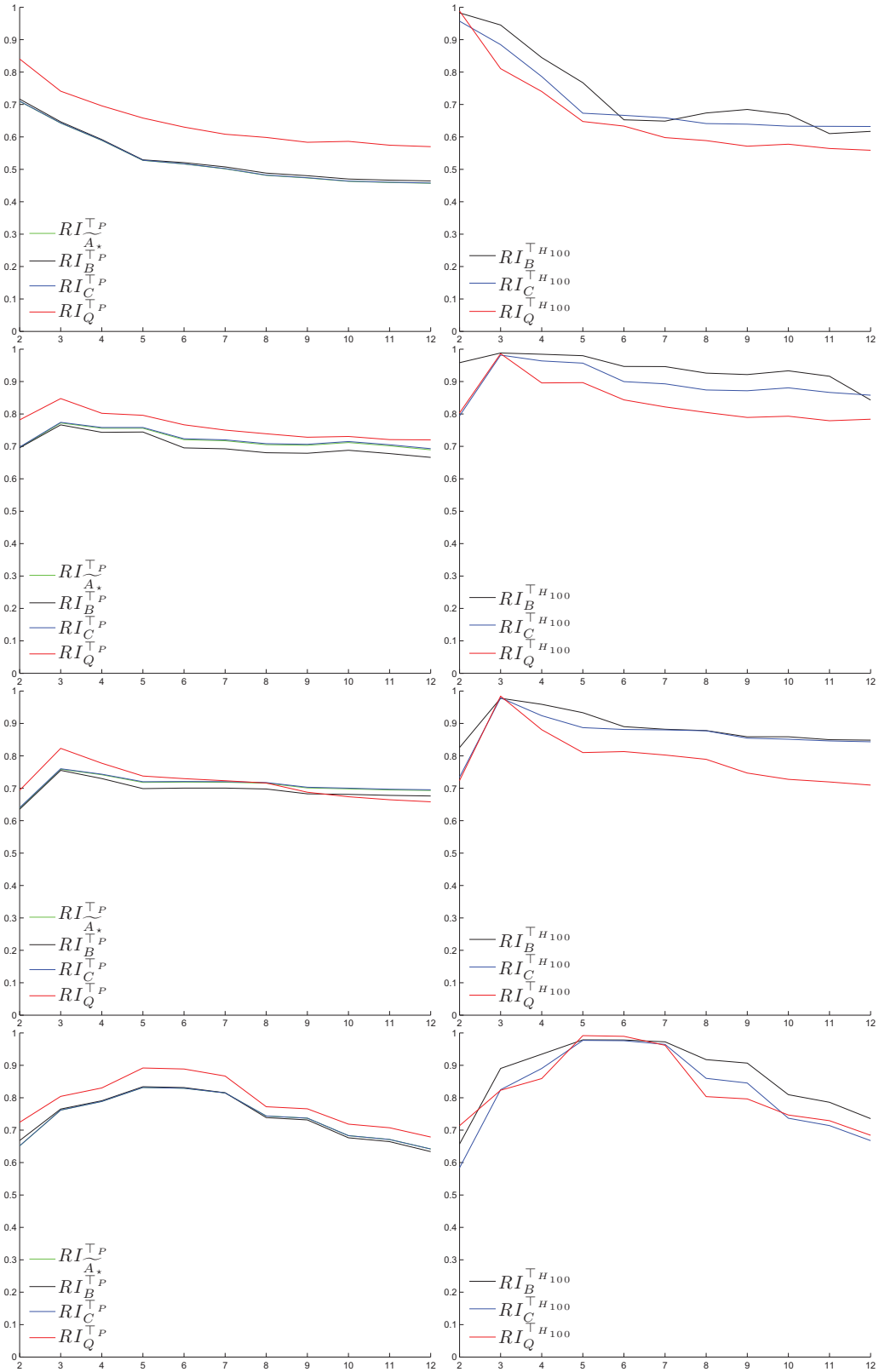



FIGURE 4.7 – Valeurs de $RI^{\top}_{\widetilde{A_\star}}$, $RI^{\top}_{C}$, $RI^{\top}_{B}$, et $RI^{\top}_{Q}$ pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

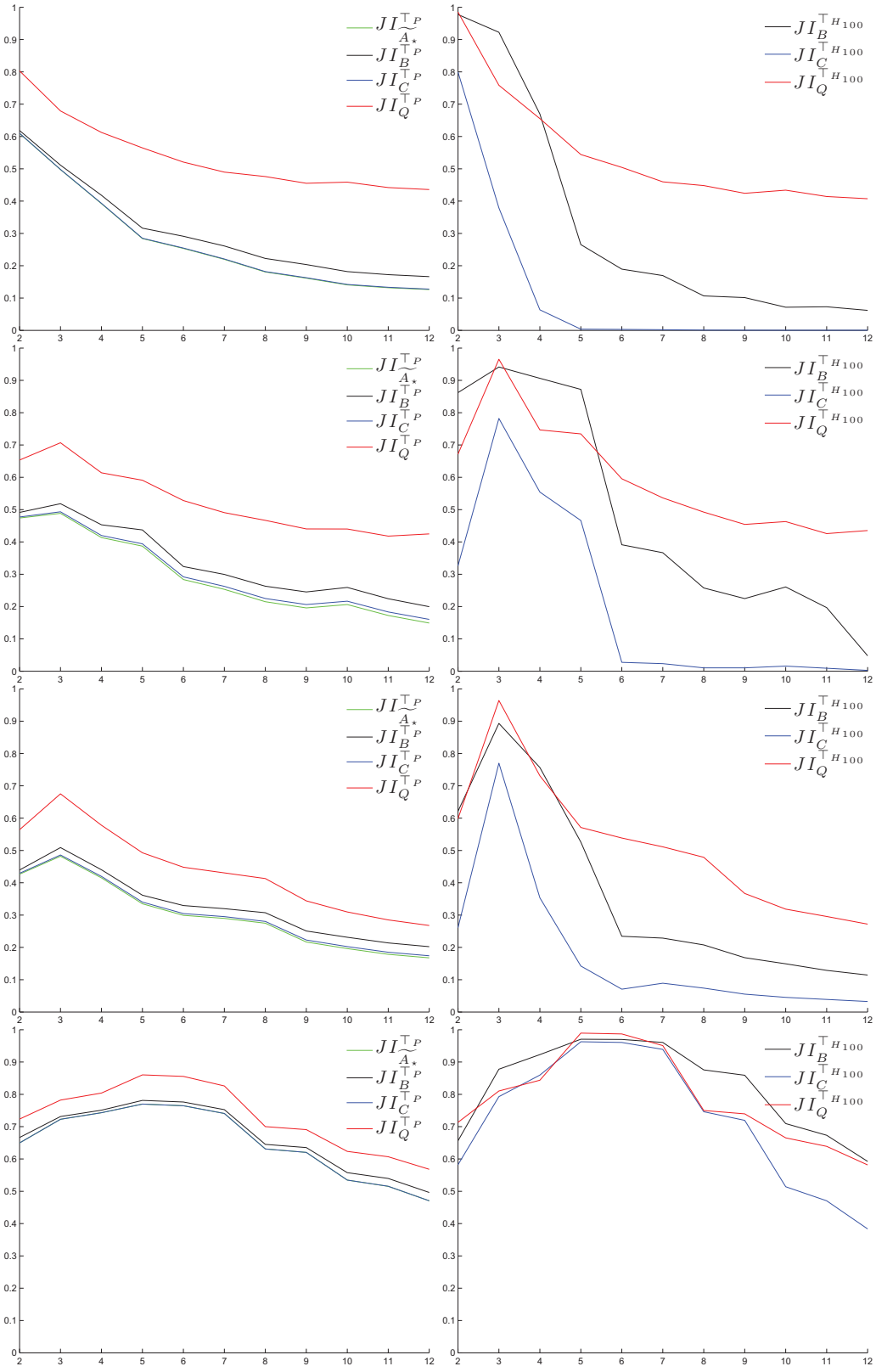



FIGURE 4.8 – Valeurs de $JI^{\top}_{\widetilde{A_\star}}$, $JI^{\top}_{C}$, $JI^{\top}_{B}$, et $JI^{\top}_{Q}$ pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

par FCM sont probablement proches de $R_{c^*}$ obtenue par le même FCM. Ce résultat est logique puisqu'aucun des indices n'est réflexif (propriété *(II-1)*) par construction. Ensuite, toutes les remarques faites à propos des résultats des expérimentations précédentes **E1** et **E3** peuvent encore être formulées ici. Nos indices de Rand présentent ainsi généralement une meilleure dynamique que les autres indices considérés, et la normalisation amoindrit l'influence de la t-norme.

### 4.3 Une mesure native orientée individus

Les expérimentations menées à la section précédente ont montré que la normalisation des matrices de coïncidence présente le défaut majeur de ne pas préserver la propriété de réflexivité *(I-1)* des indices qui en sont dérivés, au même titre que toutes les extensions d'approches orientées individus de la littérature. Cette perte s'explique aisément lorsque l'on détaille la construction mise en œuvre pour l'extension de tels indices. Elle consiste en la définition, puis l'agrégation des matrices coïncidence $\Phi_U$ et $\Phi_V$ des deux partitions à comparer. Ces deux étapes reposent sur des t-normes et des t-conormes dans le but de préserver l'acception logique de la construction stricte qu'elles se proposent d'étendre. Or, bien que l'utilisation de tels outils soit parfaitement fondée pour la première étape puisqu'ils permettent, en agrégeant deux à deux les vecteurs d'appartenance des individus aux clusters de chaque partition, de calculer pour chaque paire d'individus $\{\mathbf{x}_k, \mathbf{x}_l\}$ un degré de vérité de la proposition selon laquelle "$\mathbf{x}_k$ ET $\mathbf{x}_l$ appartiennent au même cluster", leur utilisation pour agréger les matrices de coïncidence $\Phi_U$ et $\Phi_V$ est moins fondée, comme l'illuste l'Exemple 35, cette étape devant plutôt mettre en évidence si les structures relationnelles que décrivent les deux partitions sont compatibles ou similaires.

**Exemple 35.** *Soient deux termes $\phi_{U,12} = 0.2$ et $\phi_{V,12} = 0.19$ de deux matrices de coïncidence $\Phi_U$ et $\Phi_V$ exprimant le degré selon lequel deux individus $x_1$ et $x_2$ appartiennent au même cluster dans deux partitions U et V. Avec la t-norme produit $\top_P$, la contribution de ces deux coïncidences aux termes de la matrice de contingence-paires croisant les deux partitions vaut, par (3.26) :*

$$\begin{aligned} m_{11}^{\top_P}(x_1, x_2) &= 0.2 \times 0.19 &&= 0.038, \\ m_{00}^{\top_P}(x_1, x_2) &= (1 - 0.2) \times (1 - 0.19) &&= 0.648, \end{aligned}$$

*induisant une contribution totale à l'accord entre U et V de 0.686 et de 0.038 aux indices de Rand et de Jaccard, respectivement. Les deux individus coïncidant à un niveau analogue dans chaque partition, la contribution devrait être proche de 1.*

Afin de pallier ce problème, nous proposons de définir un nouvel indice de Rand pour lequel l'agrégation des deux matrices de coïncidence est fondée, non plus sur des

normes triangulaires, mais sur une fonction de similarité (Définition 8, page 16). Cela procède de la même idée que le principe proposé par Hüllermeier et Rifqi [2009] pour leur mesure native orientée individus décrite au Chapitre 3, page 65, plus ancienne. Notre proposition présente, comme nous allons le voir par la suite, l'avantage de tirer parti à la fois de ce principe mais aussi de celui de la normalisation des matrices de coïncidence que nous avons proposée page 85, qui fait preuve de bons comportements et repose sur les bases solides de la théorie des sous-ensembles flous.

### 4.3.1 Proposition

Soit $O' = {}^t(o, o)$ un point de la diagonale $\Delta : x = y$ du carré unité $[0,1]^2$ et soient $\vec{u}$ et $\vec{n}$ deux vecteurs unité orthogonaux tels que présentés par la Figure 4.9. Chaque point $M = {}^t(x, y) \in [0,1]^2$, dont les coordonnées sont exprimées dans le repère $(O, \vec{i}, \vec{j})$, peut être exprimé par $M' = {}^t(x', y')$, dans le repère $(O', \vec{u}, \vec{n})$. Soit encore une fonction fenêtre $w$ symétrique et de moyenne nulle à valeurs dans $[0,1]$ telle que $w(0) = 1$ et une fonction profil $p$ symétrique à valeurs dans $[0,1]$ satisfaisant $p(x) > 0,\ \forall x$. Alors, la fonction $s_p^w : [0,1]^2 \to [0,1]$ définie par

$$s_p^w(x, y) = w\left(\frac{y'}{p(x')}\right) \tag{4.10}$$

est une mesure de similarité entre $x$ et $y$ au sens de la Définition 8 (page 16). Les preuves n'étant ici pas utiles, nous ne les donnons pas.

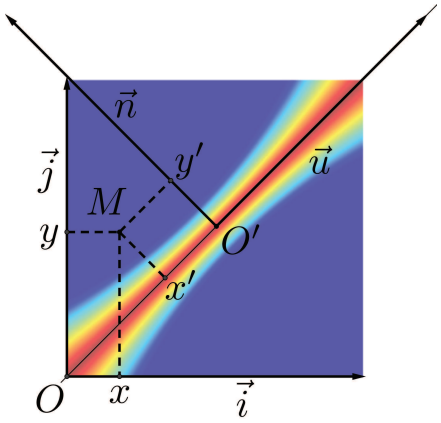



FIGURE 4.9 – Isosurface de la fonction de similarité $s_p^w$ et détails de sa construction

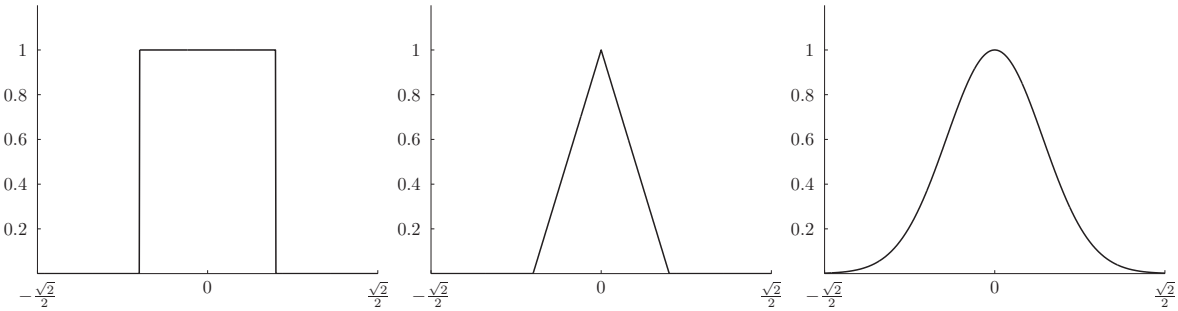
Dans l'équation (4.10), la fonction $w$ contrôle la manière dont $x$ et $y$ sont jugés similaires eût égard à leur différence. Formellement, il s'agit d'une fonction symétrique de pondération de la distance entre $M$ et $\Delta$ de telle sorte que plus $M$ s'écarte de $\Delta$, plus la valeur de $s_p^w(x, y)$ diminue. Ainsi, choisir une fonction $w$ dont la décroissance par rapport à 1 est rapide mène à une fonction de similarité drastique $s_p^w$

pour laquelle, seules les valeurs très proches de $x$ et $y$ seront considérées comme similaires. Il existe de nombreuses fonctions symétriques possibles pour $w$, et les plus communément utilisées sont reportées dans la Table 4.3. La Figure 4.10(b) présente les coupes $s_p^w$ obtenues avec les fonctions fenêtre $w_U$, $w_T$ et $w_G$, dont les isosurfaces sont représentées par la Figure 4.10(a). Ces exemples montrent comment $w$ peut radicalement transformer le comportement de la fonction de similarité $s_p^w$.

TABLE 4.3 – Fonctions fenêtre fréquentes

| | |
|---|---|
| Uniforme | $w_U(x) = 1_{(x\leq 1)}$ |
| Triangulaire | $w_T(x) = (1 - \lvert x\rvert)1_{(x\leq 1)}$ |
| Gaussien | $w_G(x) = \exp(-x^2)1_{(x\leq 1)}$ |
| Epanechnikov | $w_E(x) = (1 - x^2)1_{(x\leq 1)}$ |

(a) $w = w_U, w_T$ et $w_G$ (de gauche à droite) , $t = 0.2, r = 0, o$ est indéfini




(b) Vues en coupe de $s_p^w$ reporté plus haut le long de $x + y = 1$

FIGURE 4.10 – Les isosurfaces et les vues en coupe de quelques similarités $s_p^w$

De la même manière, la fonction de profil $p$ agit sur la similarité entre $x$ et $y$ en pondérant la différence entre ces valeurs par la distance séparant $M$ et $O'$. Cela

permet de favoriser ou non certaines gammes de valeurs spécifiques en fonction du paramètre *d'origine* choisi $o$. Nous proposons d'utiliser la fonction $p$ suivante :

$$p(x') = t + r\,x'^2 \tag{4.11}$$

où $t \in ]0, \sqrt{2}]$ et $r \in \mathbb{R}^+$ sont des paramètres définis par l'utilisateur, respectivement appelés paramètres *d'épaisseur* et *de courbure*. Il est facile de démontrer qu'avec cette fonction de profil particulière, $O'$ n'a aucune influence lorsque $r = 0$ et donc qu'aucune valeur n'est alors nécessaire pour $o$.

L'isosurface de $s_p^w$ présentée à la Figure 4.9 a pour valeurs de paramètres : $w = w_T$, une épaisseur de $t = 0.1$, une courbure de $r = 0.3$ et une origine choisie en $o = \frac{1}{2}$. On peut observer que, pour tout $M$, plus $x$ et $y$ sont proches de $\frac{1}{2}$, plus $s_p^w(x, y)$ est sélective, comme prévu. La Figure 4.11 montre également des exemples de fonctions $s_p^w$ utilisant la plupart des fonctions $w$ du Tableau 4.3 avec des valeurs spécifiques de $t$, $r$ et $o$, de telle sorte que chaque ligne de la figure mette en avant le rôle de chaque paramètre. Par exemple, la Figure 4.11(a) montre l'influence de $t$ sur la sélectivité de $s_p^w$, en fonction de sa valeur, choisie faible, moyenne et maximale, de gauche à droite.

**Définition 23.** *[Quéré et Frélicot, 2011a]*
*La concordance entre deux partitions floues $U$ et $V$ peut être mesurée par l'indice natif orienté individus $RI_{NSW}^{\top}$*[5] *défini par :*

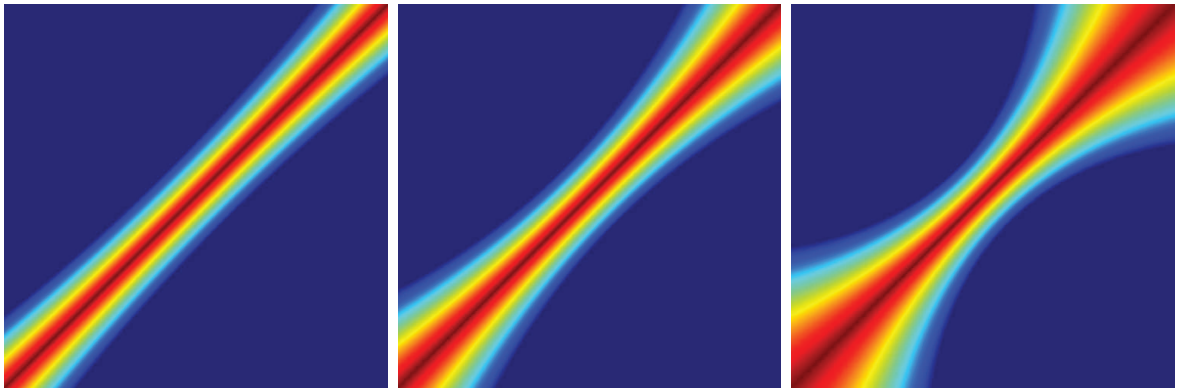
$$RI_{NSW}^{\top}(U, V) = \frac{1}{q} \sum_{k=2}^{n} \sum_{l=1}^{k-1} s_p^w(\phi_{U,kl}^{\top}, \phi_{V,kl}^{\top}) \tag{4.12}$$

*où $\phi_{U,kl}^{\top}$ et $\phi_{V,kl}^{\top}$ sont les termes des matrices de coïncidence normalisées $\Phi_U$ et $\Phi_V$ de chaque partition définies par (4.3).*

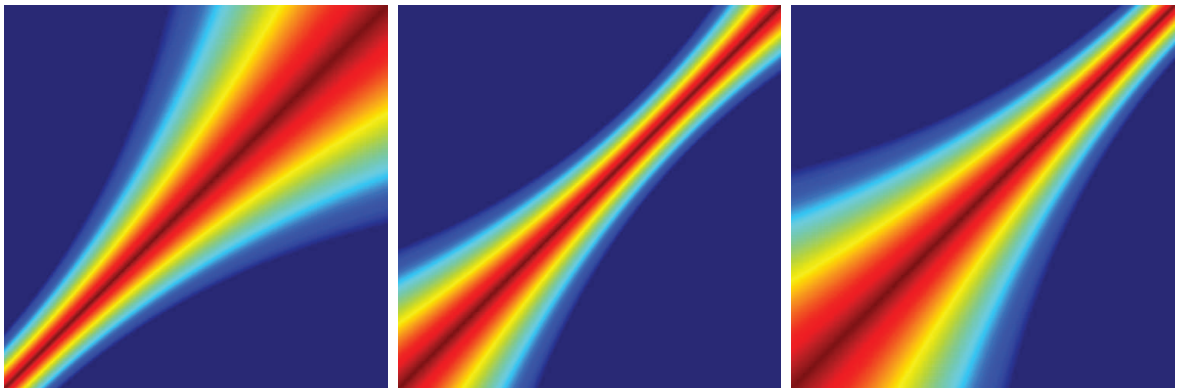
Cet indice repose sur un ensemble de cinq paramètres $\{\top, w, t, r, o\}$ définis par l'utilisateur. Quelles que soient les valeurs des paramètres utilisés pour la fonction de similarité $s_p^w$, il est facile de montrer que $RI_{NSW}^{\top}(U, V)$ se réduit à l'indice de Rand originel lorsque l'on compare deux partitions strictes, puisque dans ce cas les opérandes de $s_p^w$ sont dans $\{0, 1\}$. De plus, il atteint sa valeur maximum de 1 quand $U \equiv V$, que $U$ et $V$ soient strictes ou floues puisque $s_p^w$ est une similarité et par conséquent $s_p^w(\phi_{U_{kk}}^{\top}, \phi_{U_{ll}}^{\top}) = 1$. Il est donc réflexif (propriété (II-1)) et, pour la même raison, cet indice apporte une solution au problème illustré par l'Exemple 35. Enfin, avec la t-norme produit $\top_P$, l'invariance en échelle des coïncidences normalisées (3.28) induit, par (4.12), l'invariance en échelle de $RI_{NSW}^{\top_P}$.

5. Pour *Normalized Soft Window-based similarity Rand Index.* du fait de sa construction

(a) $w_T, t = 0.1, 0.3$ et $\frac{\sqrt{2}}{2}$ (de gauche à droite) , $r = 0, o$ est indéfini



(b) $w_T, t = 0.1, r = 0.1, 0.3$ et 0.8 (de gauche à droite) , $o = 0.5$



(c) $w_T, t = 0.1, r = 0.3, o = 0, 0.7,$ et 1 (de gauche à droite)

FIGURE 4.11 – Les isosurfaces de quelques similarités $s_p^w$

### 4.3.2 Expérimentations

Nous allons présenter et discuter ici des résultats obtenus avec notre indice $RI_{NSW}^{\top}$ pour les expérimentations décrites au début de ce Chapitre, page 79. Comme il étend

l'indice de Rand originel $RI$, il s'y ramène dans le cas de la comparaison de partitions strictes. C'est pourquoi l'expérimentation **E2** portant sur la comparaison de deux partitions strictes obtenues par clustering de jeux de données synthétiques n'est pas présentée ici. Les valeurs de $RI$ (donc de $RI^{\top}_{NSW}$) qui lui sont relatives sont présentées à la Figure 4.21, page 117.

Nous le confrontons à plusieurs propositions de la littérature revues au Chapitre 3 : la distance de transfert floue $I_{FTD_N}$ de Campello [2010], les mesures natives que sont l'indice $S^{\top}_{BH}$ de Beringer et Hüllermeier [2007] et les indices de Rand et de Jaccard $RI^{1}_{HR}$ et $JI^{1}_{HR}$ dérivés de l'approche de Hüllermeier et al. [2012]. Ces trois propositions sont décrites respectivement aux pages 64, 68 et 65. Le nombre et l'espace des paramètres utilisateur permettant de régler le nouvel indice $RI^{\top}_{NSW}$ étant vastes, nous restreignons la présentation à six paramétrages (a)-(f) donnés à la Table 4.4, choisis de manière pertinente pour illustrer le rôle de chacun des paramètres $w$, $t$, $r$ et $o$. Le choix de la t-norme n'intervient que sur les coïncidences et son incidence sur le comportement de l'indice $RI^{\top}_{NSW}$ est, par construction, sensiblement la même que pour l'indice $RI^{\top}_{Q}$. Comme elle a déjà été étudiée pour cet indice page 89, l'étude pour ce nouvel indice est reléguée en Annexe B.

TABLE 4.4 – Six paramètrages utilisés pour illustrer notre indice $RI^{\top}_{NSW}$.

| Paramétrage | $\top$ | $w$ | $t$ | $r$ | $o$ |
|---|---|---|---|---|---|
| (a) | $\top_P$ | $w_T$ | $\frac{\sqrt{2}}{2}$ | 0 | × |
| (b) | $\top_P$ | $w_T$ | 0.1 | 0 | × |
| (c) | $\top_P$ | $w_T$ | 0.1 | 0.8 | 0 |
| (d) | $\top_P$ | $w_T$ | 0.1 | 0.8 | 0.5 |
| (e) | $\top_P$ | $w_T$ | 0.1 | 0.8 | 1 |
| (f) | $\top_P$ | $w_G$ | $\frac{\sqrt{2}}{2}$ | 0 | × |

#### 4.3.2.1 Partitions non-strictes synthétiques (E1)

Nous présentons et discutons ici les résultats de l'expérimentation **E1** décrite page 80. Les mosaïques des valeurs de comparaison calculées avec les indices $I_{FTD_N}$, $S^{\top}_{BH}$, $RI^{1}_{HR}$ et $JI^{1}_{HR}$, pour les deux jeux de partitions floues $\mathcal{S}$ et possibilistes $\mathcal{S}'$ max-compatibles, sont données à la Figure 4.12. Comme les partitions de $\mathcal{S}'$ sont proportionnelles à celles de $\mathcal{S}$ et que seul le cas de la t-norme produit $\top_P$ est considéré ici, les mosaïques de notre indice $RI^{\top}_{NSW}$ ne sont données à la Figure 4.13 que pour les

partitions de $\mathcal{S}$, en vertu de la propriété d'invariance en échelle, pour cette t-norme, des coïncidences normalisées. Pour d'autres t-normes, cette propriété ne tient pas mais les observations concordent de sorte que nous préférons rejeter en Annexe B les mosaïques obtenues pour les partitions possibilistes de $\mathcal{S}'$.

La Figure 4.12 montre que les indices $I_{FTD_N}$, $S_{BH}^{\top_P}$, $RI_{HR}^{1}$ et $JI_{HR}^{1}$ sont quant à eux bien sensibles à la nature des partitions qu'ils comparent, notamment lors de la comparaison d'une partition stricte avec une partition possibiliste comme en témoigne la première ligne de chacune de leurs mosaïques. On constate ainsi que $I_{FTD_N}$ fait montre d'une faible dynamique, contrairement aux autres indices. Les résultats obtenus avec $S_{BH}^{\top_P}$ montrent par ailleurs qu'il accorde une plus grande concordance aux partitions avec un degré de flou est très important. Enfin, la diagonale blanche de leur mosaïques montre bien que chacun de ces indices est réflexif. Il en va évidemment

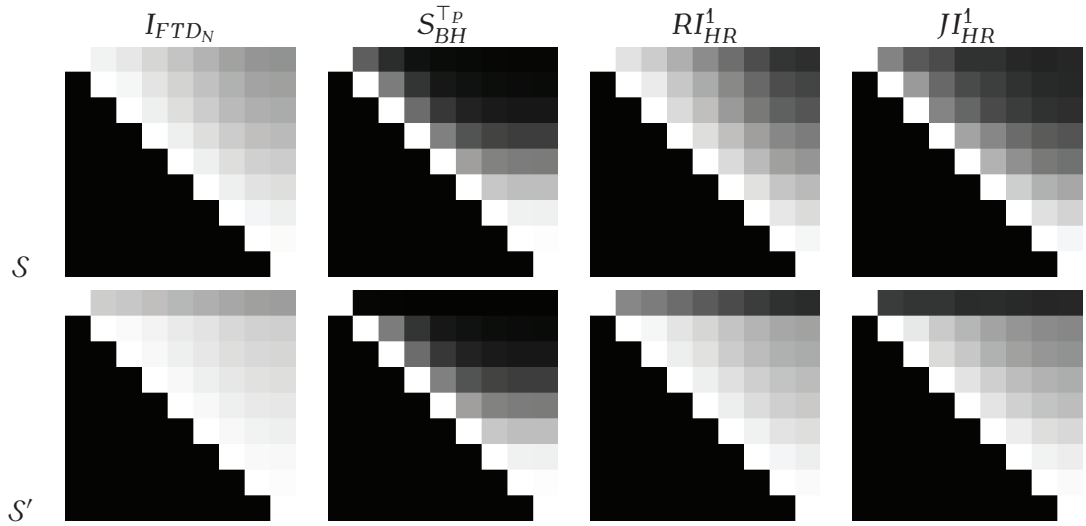



FIGURE 4.12 – Mosaïques des valeurs de comparaison obtenues avec $I_{FTD_N}$, $S_{BH}^{\top}$, $RI_{HR}^{1}$ et $JI_{HR}^{1}$. Première ligne : partitions floues de $\mathcal{S}$. Deuxième ligne : partitions possibilistes de $\mathcal{S}'$.

de même pour $RI_{NSW}^{\top}$. La Figure 4.13 montre aussi qu'en fonction du paramétrage choisi, le comportement de notre indice est très différent, ce qui lui confère une grande flexibilité. L'observation des mosaïques (a) et (b) montre ainsi l'influence du paramètre d'épaisseur $t$ : plus celui-ci est faible, plus $RI_{NSW}^{\top}$ est drastique. Les résultats obtenus avec les paramétrages (b)-(e) renseignent quant à eux sur l'influence du paramètre $r$ de courbure, en fonction de l'origine $o$ choisie. Lorsque $o$ est choisi proche de 1, une forte valeur de $r$ rend $RI_{NSW}^{\top}$ plus laxiste pour l'évaluation de la concordance entre les partitions dont le degré de flou est faible (e). Au contraire

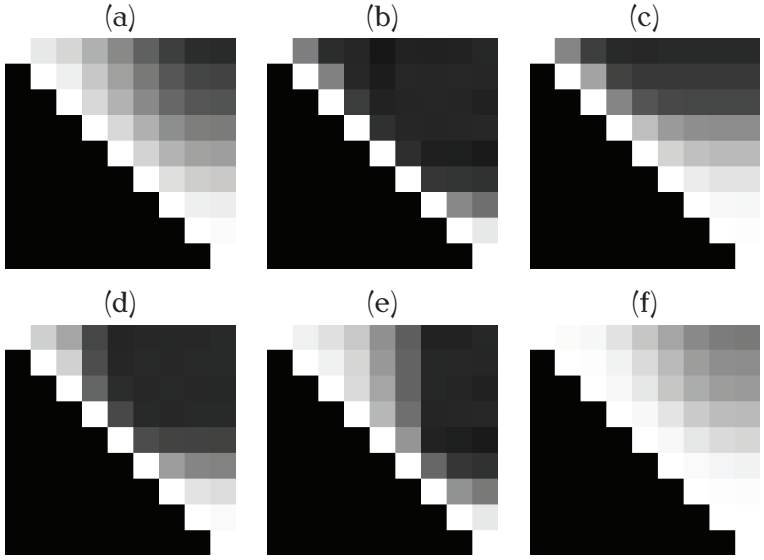



FIGURE 4.13 – Mosaïques des valeurs de comparaison obtenues avec $RI^{\top_p}_{NSW}$ selon les six paramétrages donnés par la Table 4.4 pour les partitions floues de $S$.

lorsque $o$ est choisi proche de 0, $RI^{\top}_{NSW}$ est plus conciliant avec les partitions dont le degré est fort (c). Enfin, la comparaison des mosaïques (a) et (f) souligne l'importance de la fonction de pondération $w$ choisie. Avec une fenêtre gaussienne $w_G$, $RI^{\top}_{NSW}$ présente une dynamique moins prononcée qu'avec une fenêtre triangulaire $w_T$, ce qui s'explique aisément par la forme même de ces fonctions donnée à la Figure 4.10(b), page 98.

#### 4.3.2.2 Partitions obtenues par clustering

##### 4.3.2.2.a Données synthétiques - Partitions non strictes (E3)

Nous présentons et discutons ici les résultats de l'expérimentation **E3**, décrite page 83, consacrée à la comparaison d'une partition stricte de référence avec une collection de partitions floues fournies par l'algorithme FCM et une collection de partitions possibilistes fournies par l'algorithme PCM. Ils sont donnés à la Figure 4.14, pour les indices considérés.

Un certain nombre d'observations qui ressortent de l'expérimentation précédente sont confirmées. Ainsi, on remarque que la dynamique de $I_{FTD_N}$ n'est pas exceptionnelle, contrairement à celle exhibée par $S^{\top}_{BH}$ ou les indices $RI^{1}_{HR}$ et $JI^{1}_{HR}$. Ensuite, ces indices sont très sensibles à la nature des partitions comparées, si bien que dans le cas de la comparaison des partitions possibilistes avec la partition stricte (deuxième ligne), aucun de ces indices n'atteint sa valeur maximale de 1 lorsque $\sigma = 0.1$ bien que les clusters ne présentent pas de chevauchement. Notre indice $RI^{\top}_{NSW}$ présente quant

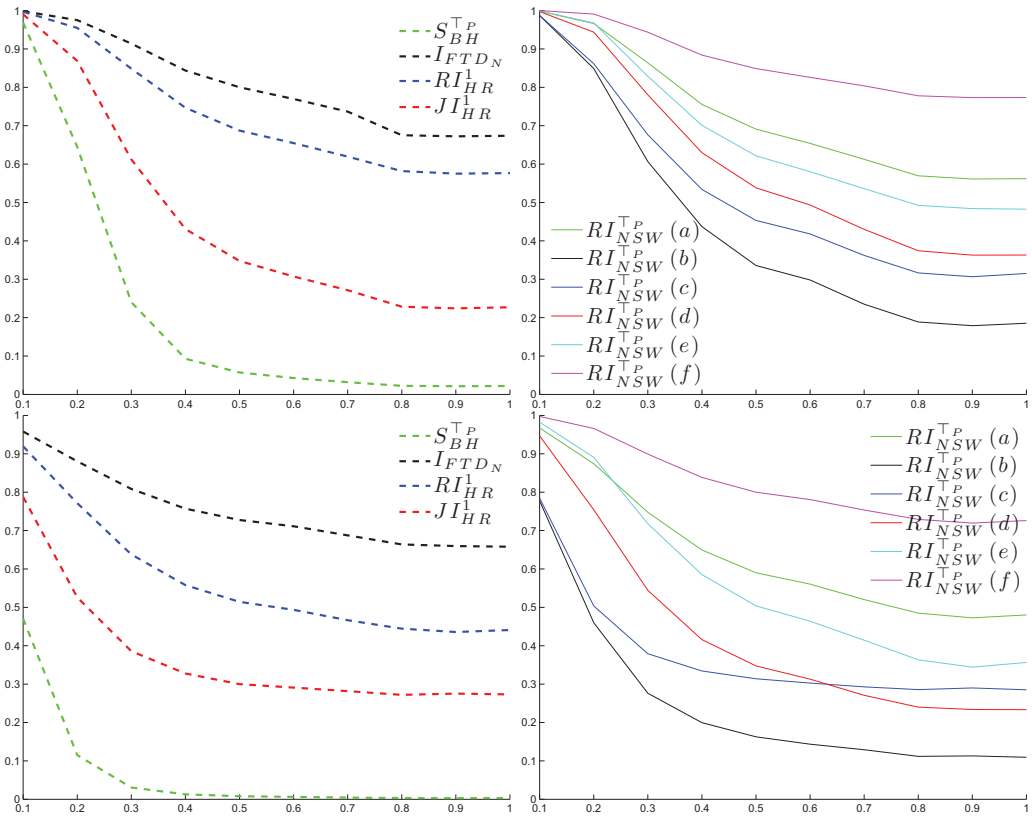



FIGURE 4.14 – Comparaison de deux collections de partitions non strictes à une partition stricte de référence $U_{k^\star}$. Première ligne : partitions floues $V_\sigma$. Deuxième ligne : partitions possibilistes $V'_\sigma$.

à lui fait des comportements variés, selon le paramétrage considéré. Tout d'abord, on remarque encore sa faible sensibilité à la nature des partitions comparées, ce qui s'explique par sa construction fondée sur la normalisation des matrices de coïncidence. Ensuite, on constate l'incidence de la fenêtre $w$ sur la dynamique de l'indice, plus forte pour $w_T$ que pour $w_G$, en comparant les courbes des paramétrages (a) et (f). De la même manière, les paramétrages (a) et (b) montrent la forte influence du paramètre d'épaisseur $t$. Plus celui-ci est faible, plus notre indice est drastique. On notera d'ailleurs la forte dynamique de la courbe associée au paramétrage (b), qui s'avère meilleure que celle exhibée par $I_{FTD_N}$, $RI_{HR}^{1}$ et $JI_{HR}^{1}$. Enfin, on remarque que le choix d'un paramètre d'origine très inférieur à 1, couplé à une forte valeur du paramètre de courbure $r$ lui confère une meilleure dynamique, en comparant les paramétrages (c), (d) et (e). À mesure que $\sigma$ augmente, et qu'avec lui augmentent le chevauchement et le degré de flou des partitions, $RI_{NSW}^{\top}$ prend des valeurs d'autant plus faibles que $o$ diminue.

#### 4.3.2.2.b Partitions non-strictes de données réelles (E4)

Nous donnons ici les résultats de l'expérimentation **E4**, décrite page 83, consacrée à la comparaison de partitions non-strictes obtenues par clustering de jeux de données réelles. Les valeurs des indices $RI^{\top}_{NSW}$, $I_{FTD_N}$, $S^{\top_P}_{BH}$, $RI^{1}_{HR}$ et $JI^{1}_{HR}$ sont donnés aux Figures 4.15 et 4.16, respectivement pour les partitions floues et possibilistes.

Toutes les remarques faites précédemment (**E1** et **E3**) peuvent encore ici être formulées, notamment celles concernant la dynamique des indices. Comme attendu puisqu'ils sont tous réflexifs, tous les indices atteignent une valeur maximale proche de 1 pour $c = c^\star$ car dans ce cas les partitions $U_{c^\star}$ (respectivement $U'_{c^\star}$) fournies par FCM (respectivement PCM) sont probablement très similaires à $R_{c^\star}$ (respectivement $R'_{c^\star}$). Nous ne commenterons pas les résultats de $I_{FTD_N}$, seule mesure non native, tant ils sont insatisfaisants quelle que soit la nature des partitions. Nous laissons le soin au lecteur de comparer les résultats de $JI^{1}_{HR}$ aux extensions de l'indice de Jaccard éprouvées à la page 95, mais on peut noter que le paramétrage (c) de notre indice $RI^{\top}_{NSW}$ lui confère un comportement approchant sur la plupart des jeux de données. De même, le paramétrage (a) rend le comportement de $RI^{\top}_{NSW}$ assez voisin de $RI^{1}_{HR}$, à comparer aux extensions dont les résultats ont été donnés à la page 94. Enfin, on peut remarquer que le paramétrage (b) le fait tendre vers l'indice $S^{\top_P}_{BH}$ dont le comportement ressemble à une fonction indicatrice tant les courbes qui sont associées à ce dernier présentent systématiquement un pic très prononcé quelle que soit la nature des partitions. Un tel comportement est très intéressant pour une mesure externe, mais le rend moins attrayant comme mesure relative avec lesquelles on cherche à établir avec une bonne granularité la concordance que partagent entre eux les éléments d'un ensemble de partitions.

## 4.4 Une mesure native orientée clusters

L'avantage principal des mesures natives de comparaison sur les extensions et/ou les généralisations est qu'on choisit les outils sur lesquels on les fonde afin de leur faire satisfaire, par construction, telle ou telle propriété. Nous présentons maintenant une proposition d'indice natif orienté clusters.

### 4.4.1 Outils préliminaires : les mesures de sparsité

La recherche d'une représentation la plus condensée possible d'un ensemble de données est un problème fondamental dans nombre de domaines. Une solution simple consiste à en trouver une représentation $\boldsymbol{y} = (y_1 \ldots y_c)$, dite parcimonieuse, pour laquelle seule une infime proportion des cœfficients $y_i$ porte à elle seule la plus

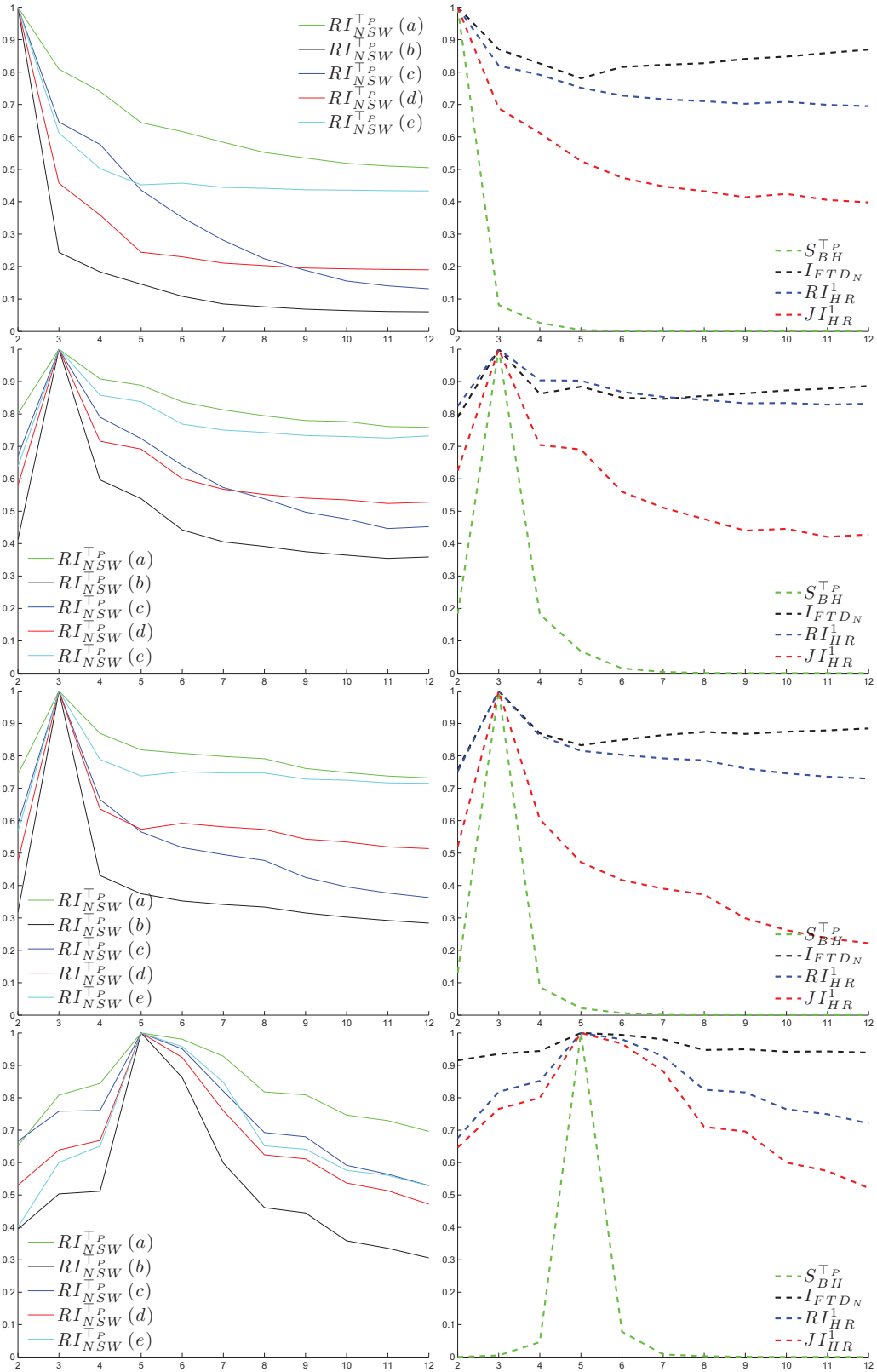



FIGURE 4.15 – Valeurs de $RI^{\top}_{NSW}$, $I_{FTD_N}$, $S^{\top_P}_{BH}$, $RI^{1}_{HR}$ et $JI^{1}_{HR}$ pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

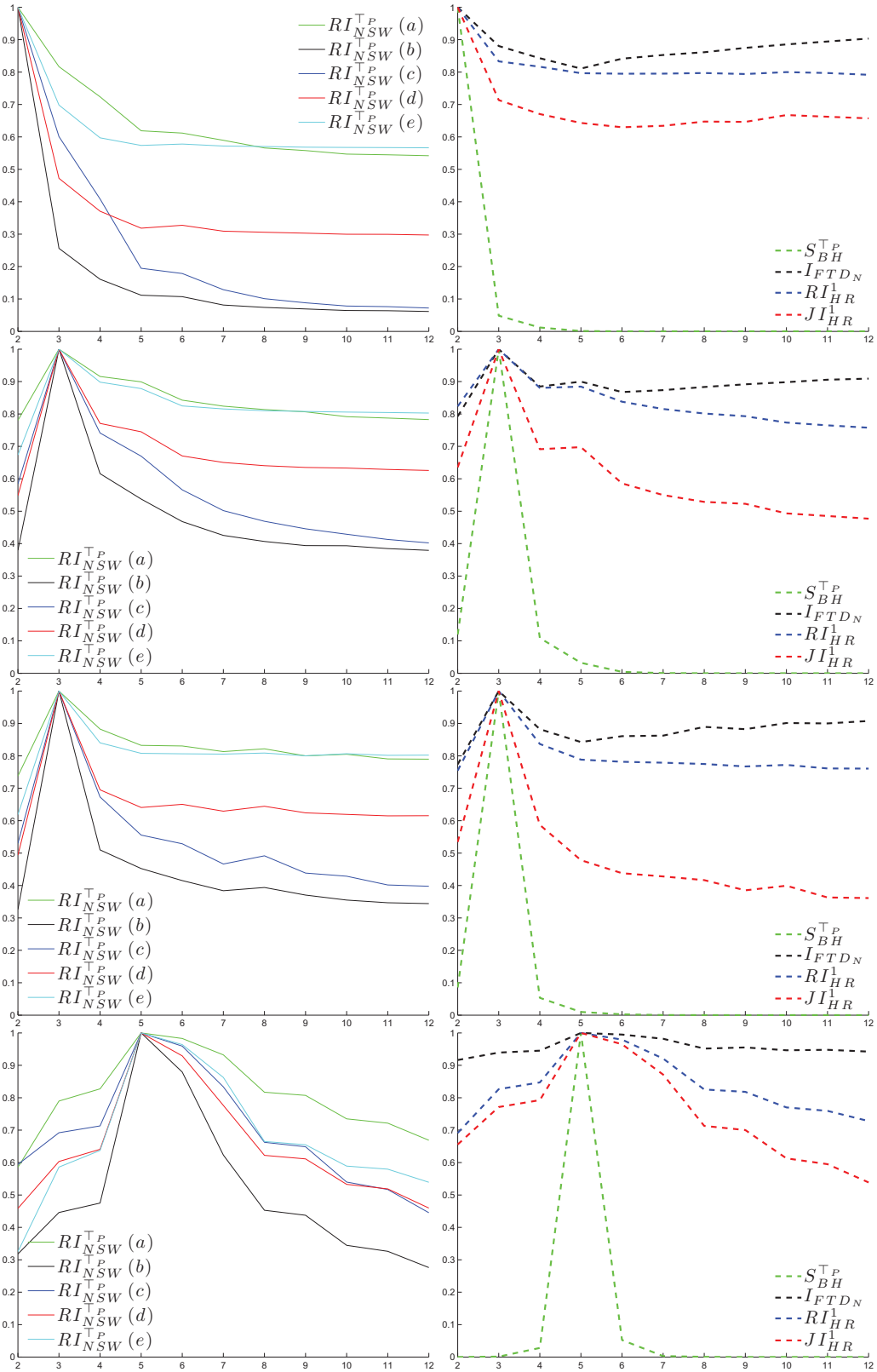



FIGURE 4.16 – Valeurs de $RI^{\top}_{NSW}$, $I_{FTD_N}$, $S^{\top_P}_{BH}$, $RI^{1}_{HR}$ et $JI^{1}_{HR}$ pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

large proportion de l'énergie de $\mathbf{y}$. Plus concrètement, un vecteur $\mathbf{y}$ est dit *sparse*, *parcimonieux* ou encore *éparse*[6] s'il est majoritairement constitué de cœfficients nuls ou proches de zéro. Ce caractère est quantifiable à l'aide d'une *mesure de sparsité* (ou de parcimonie) permettant d'évaluer à quel point $\mathbf{y}$ contient principalement des valeurs nulles ou quasi-nulles. La littérature regorge de mesures de sparsité, issues principalement de domaines tels que l'analyse et le traitement du signal [Karvanen et Cichocki, 2003].

**Définition 24.** *Une mesure de sparsité est une fonction* $S : \mathbb{R}_+^c \to \mathbb{R}$ *telle que, pour tout* $\mathbf{y} \in \mathbb{R}_+^c$ *:*

- $S\big((0\ y_2\ y_3\ \dots\ y_c)\big) > S(\mathbf{y})$,
- $S(\mathbf{y}) = S(\sigma(\mathbf{y}))$, *où* $\sigma(\mathbf{y})$ *est une permutation des cœfficients de* $\mathbf{y}$.

Une mesure de sparsité $S$ peut satisfaire plusieurs propriétés [Karvanen et Cichocki, 2003] :

(*P1*) affaiblir les forts cœfficients au profit de cœfficients plus faibles fait décroitre la sparsité :
$S\big((y_1\ \dots\ y_i - \alpha\ \dots\ y_j + \alpha\ \dots\ y_c)\big) < S(\mathbf{y})$, $\forall \alpha, y_i, y_j$ tels que $y_i > y_j$ et $0 < \alpha < \frac{y_i - y_j}{2}$,

(*P2*) un changement d'échelle ne modifie pas la sparsité :
$S(\alpha\mathbf{y}) = S(\mathbf{y})$, $\forall \alpha \in \mathbb{R}^+$,

(*P3*) ajouter la même constante à tous les cœfficients fait décroitre la sparsité :
$S(\alpha + \mathbf{y}) < S(\mathbf{y})$, $\forall \alpha \in \mathbb{R}^+$,

(*P4*) dupliquer tous les cœfficients n'influe pas sur la sparsité :
$S(\mathbf{y}) = S(\mathbf{y}||\mathbf{y}) = S(\mathbf{y}||\mathbf{y}||\mathbf{y}) = S(\mathbf{y}||\mathbf{y}||\dots||\mathbf{y})$, où $\mathbf{y}||\mathbf{z} = (y_1\ y_2\ \dots\ y_c\ z_1\ z_2\ \dots\ z_r)$ est la concaténation de deux vecteurs,

(*P5*) lorsque la valeur d'un cœfficient devient infinie, la sparsité également :
$S\big((y_1\ \dots\ y_i + \alpha\ \dots\ y_c)\big) > S\big((y_1\ \dots\ y_i\ \dots\ y_c)\big)$, $\forall \alpha > 0$,

(*P6*) l'ajout d'un cœfficent nul fait croître la sparsité :
$S(\mathbf{y}||0) > S(\mathbf{y})$.

Les principales mesures de sparsité et leur propriétés sont respectivement données aux Tables 4.5 et 4.6. Parmi ces mesures, nous ne nous intéresserons qu'à celles prenant leurs valeurs dans l'intervalle $[0, 1]$ et satisfaisant la propriété (*P5*) pour des raisons qui seront exposées plus loin, page 112. Sont ainsi retenues le Kurtosis $\kappa_4$, la mesure de Hoyer $H$, la $pq$-moyenne $E_{p,q}$ dont la précédente en est, à un facteur près, le cas particulier $(p = 2, q = 4)$, et le critère de Gini $G$, qui vérifient toutes :

6. Ces deux derniers termes étant très peu utilisés dans la littérature, nous dirons que $\mathbf{y}$ est *sparse*.

- $S\big((y_1\ 0\ 0\ ...\ 0)\big) = 1,\ \forall y_1 \in \mathbb{R}_+$,
- $S\big((y_1\ y_1\ y_1\ ...\ y_1)\big) = 0,\ \forall y_1 \in \mathbb{R}_+$.

Afin de mettre en lumière le comportement de chacune de ces quatre mesures, nous avons réalisé deux petites expérimentations. La première consiste en l'étude de la vitesse de convergence en $\alpha$ pour la propriété (P5). Pour ce faire, nous considérons un vecteur $\mathbf{y}_\alpha = (\alpha\ 1\ 1\ ...\ 1) \in \mathbb{R}^{15}$, pour lequel nous faisons varier $\alpha \in [0, 150]$. Pour chaque mesure de sparsité $S$, la courbe $S(\mathbf{y}_\alpha)$ résultante est donnée à la Figure 4.17. C'est le Kurtosis qui converge le plus rapidement puisque sa valeur dépasse 0.9 dès que $\alpha > 20$, tandis qu'à l'opposé le critère de Gini $G$ n'atteint pas 0.8 tant que $\alpha < 60$. Ceci indique que le Kurtosis réagit bien plus fortement que toutes les autres mesures à l'apparition d'un cœfficient se détachant ostensiblement. La seconde expérimenta-

TABLE 4.5 – Principales mesures de sparsité $S(\mathbf{y})$ [Hurley et Rickard, 2009]

| Mesure | $S(\mathbf{y})$ | Intervalle de sortie |
|---|---|---|
| Norme $-l_0$ | $card\big(\ \{y_i \vert y_i > 0\}\ \big)$ | $\mathbb{Z}$ |
| Norme $-l_0^\epsilon$ | $card\big(\ \{y_i \vert y_i > \epsilon\}\ \big)$ | $\mathbb{Z}$ |
| Norme $-l_1$ | $-\sum_{j=1}^{c} y_j$ | $\mathbb{R}$ |
| Norme $-l_p$, $0 < p < 1$ | $-\left(\sum_{j=1}^{c} y_j^p\right)^{1/p}$ | $\mathbb{R}$ |
| Rapport $\frac{l_2}{l_1}$ | $\frac{\sqrt{\sum_{j=1}^{c} y_j^2}}{\sum_{j=1}^{c} y_j}$ | $\mathbb{R}$ |
| Tangente hyperbolique $-th_{a,b}$ | $-\sum_{j=1}^{c} tanh\left(a^b y_j^b\right)$ | $]-1, 1[$ |
| Logarithme $-log$ | $-\sum_{j=1}^{c} log(1 + y_j^2)$ | $\mathbb{R}$ |
| Kurtosis $\kappa_4$ | $\frac{\frac{\sum_{j=1}^{c} y_j^4}{\left(\sum_{j=1}^{c} y_j^2\right)^2} - \frac{1}{c}}{1 - \frac{1}{c}}$ | $[0, 1]$ |
| Entropie normalisée de Shannon $\widehat{H}_s$ | $-\sum_{j=1}^{c} \frac{y_j^2}{l_2(\mathbf{y})^2} log\left(\frac{y_j^2}{l_2(\mathbf{y})^2}\right)$ | $\mathbb{R}$ |
| Mesure de Hoyer $H$ | $\frac{\sqrt{c} - \frac{\sum_{j=1}^{c} y_j}{\sqrt{\sum_{j=1}^{c} y_j^2}}}{\sqrt{c} - 1}$ | $[0, 1]$ |
| $pq$-moyenne $E_{p,q}$, $p < q$ | $\frac{1 - \left(\frac{1}{c}\sum_{j=1}^{c} y_j^p\right)^{\frac{1}{p}} \left(\frac{1}{c}\sum_{j=1}^{c} y_j^q\right)^{-\frac{1}{q}}}{1 - c^{\frac{p-q}{pq}}}$ | $[0, 1]$ |
| Critère de Gini $G$ | $\frac{1 - 2\sum_{k=1}^{c} \frac{y_{(k)}}{l_0(\mathbf{y})}\left(\frac{c - k + \frac{1}{2}}{c}\right)}{1 - \frac{1}{c}}$ avec $\mathbf{y}$ trié tel que : $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(c)}$ | $[0, 1]$ |

TABLE 4.6 – Propriétés des mesures de sparsité $S(\mathbf{y})$ de la Table 4.5

| Mesure | (P1) | (P2) | (P3) | (P4) | (P5) | (P6) |
|---|---|---|---|---|---|---|
| Norme $-l_0$ | | × | | | | × |
| Norme $-l_0^\epsilon$ | | | | | | × |
| Norme $-l_1$ | | | × | | | |
| Norme $-l_p$, $0 < p < 1$ | × | | × | | | |
| Rapport $\frac{l_2}{l_1}$ | × | × | | | × | |
| Tangente hyperbolique $-th_{a,b}$ | × | | × | | | |
| Logarithme $-log$ | | | × | | | |
| Kurtosis $\kappa_4$ | | × | × | | × | × |
| Entropie normalisée de Shannon $\widehat{H}_s$ | | | | | | |
| Mesure de Hoyer $H$ | × | × | × | | × | × |
| $pq$-moyenne $E_{p,q}$, $p < q$ | $p \leq 1, q > 1$ | × | × | × | × | × |
| Critère de Gini $G$ | × | × | × | × | × | × |

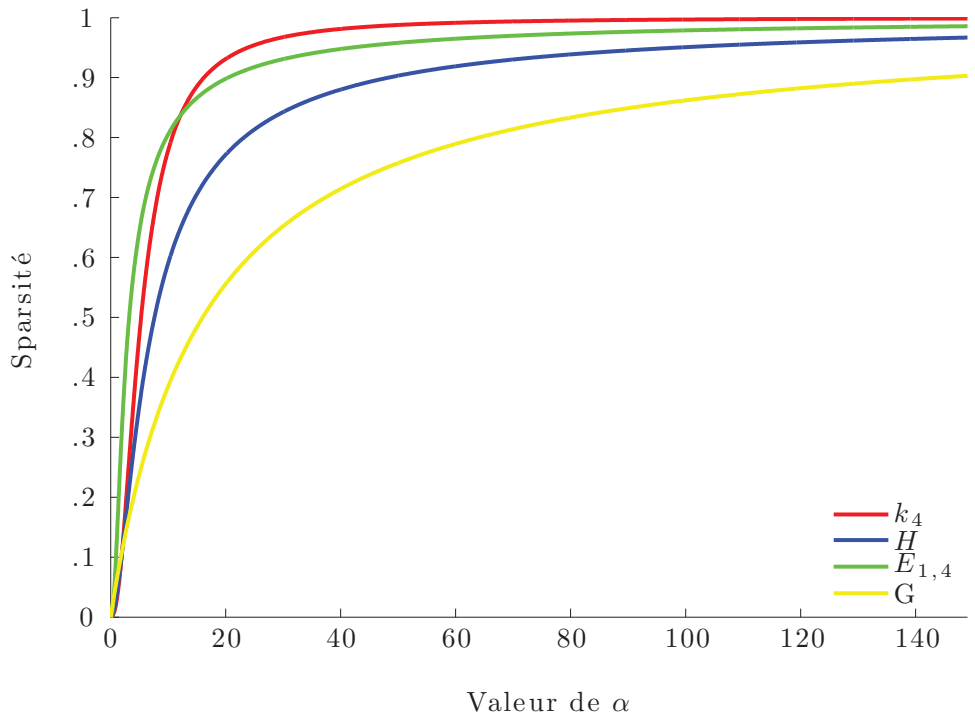



FIGURE 4.17 – Convergence en $\alpha$ de quatre mesures de sparsité satisfaisant (P5)

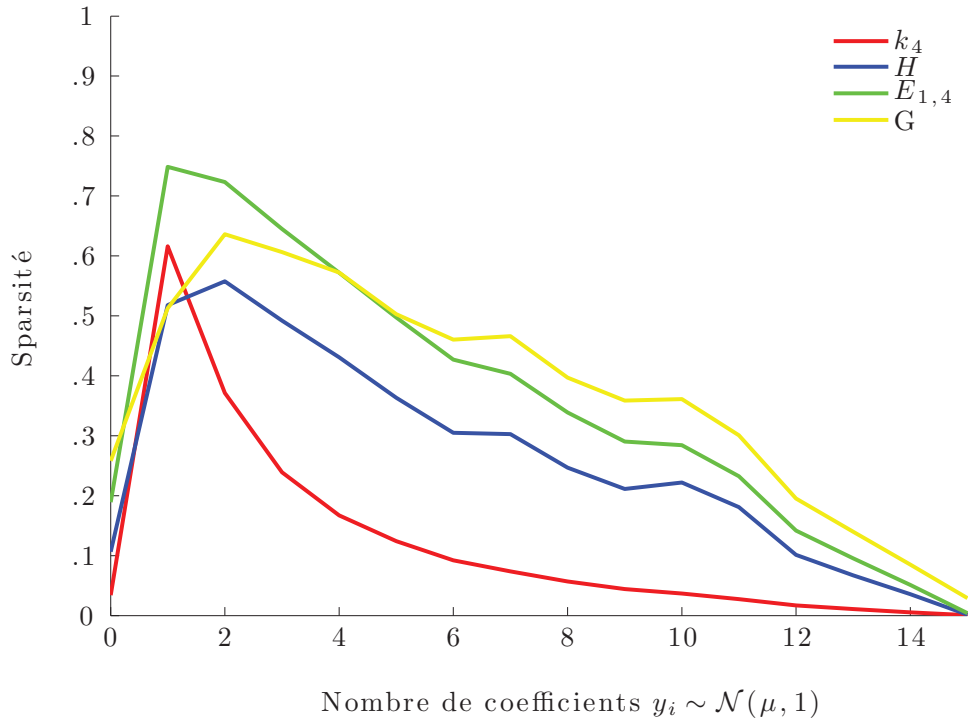



FIGURE 4.18 – Pouvoir discriminant de quatre mesures de sparsité

tion vise à étudier comment les mesures de sparsité retenues se comportent à mesure qu'un nombre $N < c$ des cœfficients forts augmente. Pour cela, nous avons généré des vecteurs $\mathbf{y}_N$ composés de $N$ cœfficients indépendants et identiquement distribués selon une loi normale $\mathcal{N}(\mu = 20, \sigma = 1)$ et de $M = c - N$ cœfficients $\epsilon$ tirés selon une loi uniforme $\mathcal{U}(0,5)$. Pour chaque mesure de sparsité $\mathcal{S}$ considérée, la courbe $\mathcal{S}(\mathbf{y}_N)$ est tracée à la la Figure 4.18 en fonction de $N$ variant de 1 à 15. Tout d'abord, on constate que chaque mesure présente un pic autour de $N = 1$, ce qui résulte de la propriété (*P5*). À cause du bruit uniforme $\epsilon$ présent dans $\mathbf{y}_N$, aucune d'entre elles n'atteint toutefois son maximum de 1. Ensuite, on remarque que le Kurtosis se différencie encore une fois des autres mesures par une décroissance plus rapide à mesure que $N$ s'éloigne que 1. Il présente donc un fort pouvoir discriminant eut égard à la propriété (*P5*), en distinguant fortement le vecteur $\mathbf{y}_1$ pour lequel un seul cœfficient est fort. La mesure de Hoyer et la $pq$-moyenne ont un comportement analogue, expliqué par le fait que la première est un cas particulier de la seconde. Enfin, le critère de Gini paraît très sensible au bruit, et semble posséder un faible pouvoir discriminant. Ces observations justifieront certains choix faits pour les expériementations menées à la page 114.

### 4.4.2 Proposition

Afin d'expliquer les principes sur lesquels repose l'indice que nous proposons maintenant, conçu pour comparer des partitions non strictes, considérons les deux partitions strictes de l'Exemple 4 que nous avons largement utilisées tout au long de ce mémoire, dès la page 9 :

$$U_h = \begin{array}{c} \begin{matrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_4 & \mathbf{u}_5 \end{matrix} \\ \left( \begin{matrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{matrix} \right) \end{array} \begin{matrix} U_h^1 \\ U_h^2 \end{matrix} \qquad V_h = \begin{array}{c} \begin{matrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 \end{matrix} \\ \left( \begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{matrix} \right) \end{array} \begin{matrix} V_h^1 \\ V_h^2 \\ V_h^3 \end{matrix}$$

Ces deux partitions partagent en commun de regrouper au sein de deux clusters les individus $\mathbf{x}_3$ et $\mathbf{x}_4$ et les individus $\mathbf{x}_2$ et $\mathbf{x}_5$. Elles ne diffèrent que par $\mathbf{x}_1$, qui est regroupé avec $\mathbf{x}_3$ et $\mathbf{x}_4$ dans $U_h$ tandis qu'il est mis de côté dans son propre cluster dans $V_h$. Cette différence apparaît clairement dans la matrice de contingence croisant les clusters de $U_h$ et $V_h$, donnée à l'Exemple 13, page 29 : $N(U_h, V_h) = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. En effet, la première ligne $\mathbf{n}_1 = (1\ 2\ 0)$ de $N(U_h, V_h)$ montre que les trois individus qui composent $U_h^1$ sont répartis dans les clusters $V_h^1$ et $V_h^2$ de $V_h$, et traduit donc un désaccord entre les partitions. Au contraire, la seconde ligne $\mathbf{n}_3 = (0\ 0\ 2)$ et la première colonne ${}^t\mathbf{n}_3 = (0\ 2)$ témoignent du fait que $U_h^2$ et $V_h^3$ sont identiquement composés des mêmes éléments. De la même manière, si l'on observe les matrices de contingence $N(U_h, U_h) = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$ et $N(V_h, V_h) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ qui croisent chaque partition avec elle-même, on remarque dans ce cas que chacune de ces matrices ne présente logiquement qu'un seul élément non nul sur chaque ligne et chaque colonne. Ainsi, notre proposition repose sur la constatation suivante : plus les partitions sont concordantes, plus les lignes et colonnes de leur matrice de contingence sont sparses. Cette idée est également valable lorsque l'on croise deux partitions floues $U_f$ et $V_f$. En effet, le produit scalaire entre $U_f^i$ et $V_f^j$ induit par (2.20) sera élevé si et seulement si les deux clusters sont similaires et ne sont pas trop flous, *i.e.* s'ils ont une certaine quantité $n_s < n$ d'appartenance en commun proches de 1, de telle sorte que $U_f^i \, {}^t V_f^j \to n_s$. En considérant l'ensemble des clusters de $U_f$ et $V_f$, la valeur de chacun des éléments de $N(U_f, V_f)$ sera donc forte lorsque les clusters flous correspondants présenteront une forte ressemblance et sera faible pour ceux qui diffèrent considérablement, de telle sorte que dans ce cas encore la sparsité des lignes et colonnes de $N(U_f, V_f)$ variera en fonction de la proximité des partitions comparées.

Concrètement, le nouvel indice est construit de la manière suivante. Étant donnée une mesure de sparsité $\mathcal{S}$ à valeur dans $[0,1]$, parmi celles retenues à la section précédente[7], on peut définir les deux ensembles suivants, composés des sparsités de chaque ligne et colonne de la matrice de contingence $N(U,V)$ de deux partitions $U$ et $V$ de $\mathbb{M}_{pcn}$ :

$$\mathcal{R}_{\mathcal{S}} = \{\mathcal{S}(\mathbf{n_1}), ..., \mathcal{S}(\mathbf{n_r})\} \text{ et } \mathcal{C}_{\mathcal{S}} = \{\mathcal{S}({}^{t}\mathbf{n_1}), ..., \mathcal{S}({}^{t}\mathbf{n_c})\} \tag{4.13}$$

Chacun de ces ensembles $\mathcal{R}_{\mathcal{S}}$ et $\mathcal{C}_{\mathcal{S}}$ est ensuite agrégé à l'aide d'une fonction d'agrégation $\mathcal{A}$ afin d'obtenir deux valeurs $\mathcal{A}(\mathcal{R}_{\mathcal{S}})$ et $\mathcal{A}(\mathcal{C}_{\mathcal{S}})$ représentatives de la sparsité des lignes et des colonnes, respectivement. Il existe de nombreuses familles de fonctions d'agrégation et le lecteur pourra se référer à [Grabisch et al., 2009] pour une monographie récente. Nous imposons aux fonctions d'agrégation de prendre leurs valeurs dans $[0,1]$, car leur sorties seront des entrées d'une implication résiduelle floue $\mathcal{I}_{\top}$, définie par (3.15) page 51.

**Définition 25.** *Une fonction d'agrégation est une fonction c-aire $\mathcal{A} : [0,1]^c \to [0,1]$ satisfaisant pour tout $\mathbf{a}, \mathbf{b} \in [0,1]^c$ les propriétés suivantes :*

- $\mathcal{A}(0,0,...,0) = 0$ *et* $\mathcal{A}(1,1,...,1) = 1$ *(conditions aux bornes),*
- $a_1 \leq b_1,\ a_2 \leq b_2, ..., a_c \leq b_c \Rightarrow \mathcal{A}(a_1, a_2, ..., a_c) \leq \mathcal{A}(b_1, b_2, ..., b_c)$ *(monotonie).*

Dans les expérimentations menées à partir de la page 114, nous nous limiterons à la moyenne arithmétique $\mathcal{A}_{\Sigma}(\mathbf{a}) = \frac{1}{c}\sum_{i=1}^{c} a_i$ qui est à valeurs dans $[0,1]$ lorsque $a \in [0,1]^c$.

**Définition 26.** *[Quéré et Frélicot, 2012]*
*La concordance entre deux partitions possibilistes U et V peut être mesurée par l'indice natif orienté clusters QF défini, à partir des sparsités des lignes ($\mathcal{R}_{\mathcal{S}}$) et des colonnes ($\mathcal{C}_{\mathcal{S}}$) de leur matrice de contingence $N(U,V)$, par :*

$$QF_{(\mathcal{S},\mathcal{A},\mathcal{I}_{\top})}(U,V) = min\Big(\mathcal{I}_{\top}\big(\mathcal{A}(\mathcal{R}_{\mathcal{S}}), \mathcal{A}(\mathcal{C}_{\mathcal{S}})\big), \mathcal{I}_{\top}\big(\mathcal{A}(\mathcal{C}_{\mathcal{S}}), \mathcal{A}(\mathcal{R}_{\mathcal{S}})\big)\Big), \tag{4.14}$$

*où $\mathcal{I}_{\top}$ est une implication résiduelle floue (définie par (3.15), page 51) satisfaisant le principe de confinement (IF-7) (page 50), $\mathcal{A}$ est une fonction d'agrégation (Définition 25) et $\mathcal{S}$ est une mesure de sparsité vérifiant la propriété (P5).*

On peut interpréter cet indice comme la valeur de vérité de la proposition floue suivante : "si la sparsité en ligne implique la sparsité en colonne ET si la sparsité

7. La mesure $\mathcal{S}$ doit satisfaire la propriété (P5), afin de présenter une forte valeur pour des lignes ou colonnes de la matrice de contingence composées d'un seul cœfficient non nul ou proche de zéro.

en colonne implique la sparsité en ligne, alors les deux partitions concordent". Par construction, ce nouvel indice est symétrique, prend ses valeurs dans $[0,1]$ et atteint son maximum lorsque $N(U,V)$ n'est composée que d'un seul cœfficient non nul par ligne et par colonne. L'implication $\mathcal{I}_\top$ doit satisfaire le principe de confinement *(IF-7)*, page 50, afin que l'indice soit réflexif (propriété *(I-1)*), de sorte que $QF_{(\mathcal{S},\mathcal{A},\mathcal{I}_\top)}(U,V) = 1$ lorsque $U \equiv V$. Enfin, il est important de noter que cet indice présente une complexité asymptotique en temps de l'ordre de $\mathcal{O}(n)$, le plaçant parmi les indices les plus rapides répertoriés à la Table 3.14, page 75, comme l'indice de Anderson et al. [2010] et la distance de transfert floue proposée par Campello [2010], devant beaucoup d'autres comme les indices de Hüllermeier et al. [2012].

### 4.4.3 Expérimentations

Encore une fois, nous allons étudier le comportement de notre indice $QF_{(\mathcal{S},\mathcal{A},\mathcal{I}_\top)}$ à partir des résultats obtenus pour la série d'expérimentations décrites au début de ce chapitre, page 79. Ces résultats sont confrontés à ceux obtenus avec plusieurs indices de la littérature : la distance de transfert floue $I_{FTD_N}$ de Campello [2010], l'indice natif $S^\top_{BH}$ de Beringer et Hüllermeier [2007] et les indices natifs de Rand et de Jaccard $RI^1_{HR}$ et $JI^1_{HR}$ dérivés de l'approche de Hüllermeier et al. [2012], auxquels nous avons déjà comparé notre précédente proposition.

L'étude est limitée à quatre paramétrages, combinant deux mesures de sparsité et deux implications résiduelles floues, pour la seule mesure d'agrégation qu'est la moyenne arithmétique $\mathcal{A}_\Sigma$. Parmi les mesures de sparsité décrites page 110, nous avons retenu la mesure de Hoyer $H$ et le Kurtosis $\kappa_4$ qui satisfont les contraintes définies par notre proposition et présentent des comportements à la fois satisfaisants et sensiblement différents, comme nous l'avons montré à la page 105 du présent chapitre. De la même manière, les implications de Gödel $\mathcal{I}_M$ et de Hamacher $\mathcal{I}_{H_0}$, dont la définition a été donnée à la Table 3.4 du Chapitre 3 (page 51) ont été retenues pour leur contraste. L'indice de $S^\top_{BH}$ est quant à lui calculé avec la t-norme produit $\top_P$.

#### 4.4.3.1 Partitions non-strictes synthétiques (E1)

Cette expérimentation **E1**, décrite page 80, consiste à comparer neuf partitions floues $(S)$ entre elles, et neuf partitions possibilistes $(S')$. Les mosaïques résultant des différents paramétrages de notre indice $QF_{(\mathcal{S},\mathcal{A},\mathcal{I}_\top)}$ sont donnés à la Figure 4.19, celles des indices $I_{FTD_N}$, $S^\top_{BH}$, $RI^1_{HR}$ et $JI^1_{HR}$, déjà présentés page 116, sont rappelés à la Figure 4.20 pour le confort de lecture.

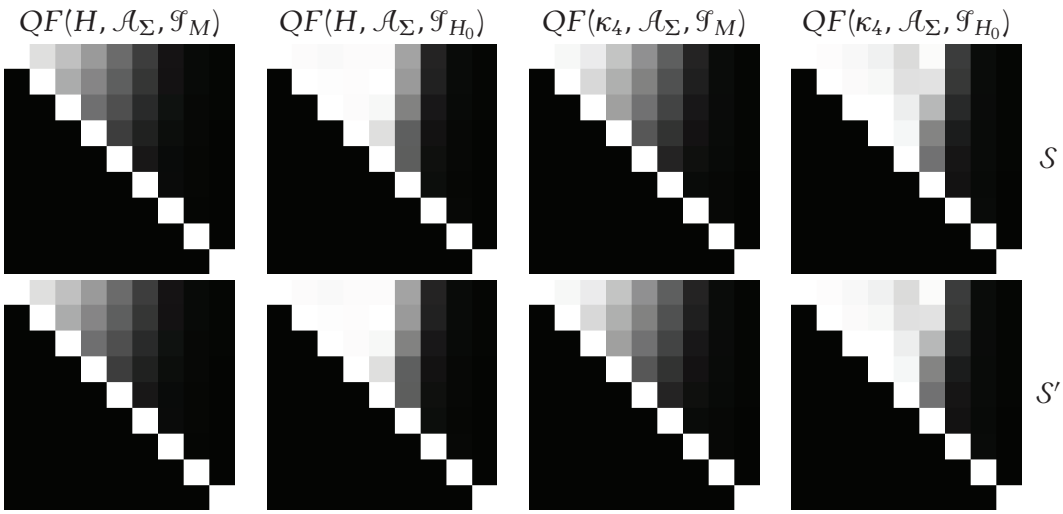



FIGURE 4.19 – Mosaïques des valeurs de comparaison obtenues avec $QF$ selon quatre paramétrages. Première ligne : partitions floues de $S$. Deuxième ligne : partitions possibilistes de $S'$.

À l'observation des mosaïques de la Figure 4.19, plusieurs remarques viennent à l'esprit. Tout d'abord, notre proposition s'avère peu sensible à la nature des partitions puisque les mosaïques correspondant aux partitions floues ($S$) sont similaires à celles associées aux partitions possibilistes ($S'$). Il s'agit d'une conséquence de la propriété d'invariance en échelle (*P2*) satisfaite par les mesures de sparsité considérées. Ensuite, la diagonale blanche exhibée par chacune des mosaïques montre que notre indice $QF_{(S,\mathcal{A}_\Sigma,\mathcal{I}_\top)}$ est bien réflexif, et qu'il atteint sa valeur maximale de 1 lors de la comparaison d'une partition avec elle-même. On remarque aussi que, contrairement aux autres indices (Figure 4.20), notre indice n'établit aucun accord entre les partitions dont le degré de flou est très fort (partie inférieure droite des mosaïques). Ce comportement s'explique par le fait que de telles partitions sont composées d'éléments qui ont tous des valeurs très proches. Lorsqu'on les croise, les matrices de contingence induites possèdent logiquement une très faible sparsité. Par ailleurs, on constate que l'implication de Hamacher $\mathcal{I}_{H_0}$ amène $QF_{(S,\mathcal{A}_\Sigma,\mathcal{I}_\top)}$ à considérer comme très concordantes les partitions présentant un faible degré de flou (partie supérieure gauche des mosaïques). A contrario, avec l'implication résiduelle de Gödel $\mathcal{I}_M$, il fait apparaitre une transition plus douce entre les partitions dont le degré de flou est faible et celles pour lesquelles il est plus fort. Enfin, nous nous retiendrons d'émettre ici tout commentaire à propos des différences induites par chacune des mesures $H$ et $\kappa_4$ utilisées pour le calcul de $QF_{(S,\mathcal{A}_\Sigma,\mathcal{I}_\top)}$, celles-ci étant bien mieux explicitées par les résultats des expérimentations suivantes.

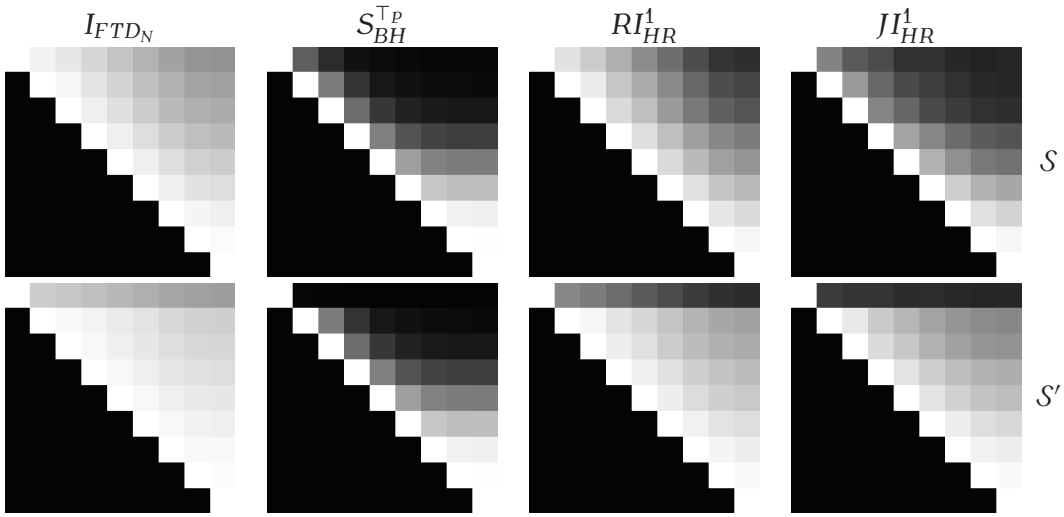



FIGURE 4.20 – Mosaïques des valeurs de comparaison obtenues avec $I_{FTD_N}$, $S^{\top}_{BH}$, $RI^1_{HR}$ et $JI^1_{HR}$. Première ligne : partitions floues de $\mathcal{S}$. Deuxième ligne : partitions possibilistes de $\mathcal{S}'$.

#### 4.4.3.2 Partitions obtenues par clustering

##### 4.4.3.2.a Données synthétiques – Partitions strictes (E2)

Puisque notre indice $QF_{(\mathcal{S},\mathcal{A},\mathcal{F}_\top)}$ n'est ni une extension, ni une généralisation d'aucun indice strict, mais parce que sa construction permet de comparer entre autres des partitions strictes, nous présentons ici les résultats de l'expérimentation **E2**, décrite page 81, dédiée à la comparaison d'une collection de partitions strictes $V_k$ ($k = 2, ..., 12$) obtenues par $k - Moyennes$ avec une partition stricte $U_{k^\star}$ de référence. Notre proposition est comparée à des indices stricts parmi les plus populaires de la littérature : les indices de Rand $RI$ et de Jaccard $JI$ stricts, et la distance de transfert $TD$, revus au Chapitre 2 aux pages 21 et 33. Les résultats obtenus avec ces trois indices ainsi que notre indice $QF$ calculé pour les quatre paramétrages retenus sont donnés par la Figure 4.21.

Tous les indices considérés atteignent bien leur maximum pour la partition $V_{k^\star}$, comme attendu, mais avec des dynamiques assez différentes. On constate que celle de l'indice de Rand historique est assez faible, contrairement à celle de l'indice de Jaccard $JI$ auquel il fait pourtant souvent de l'ombre. Cette différence est due à la prise en compte ($RI$) ou non ($JI$) du terme d'accord $m_{00}$ de la matrice de contigence-paires, comme nous avons pu en discuter plus tôt au Chapitre 2, page 22. La distance distance de transfert $TD$ présente quant à elle une bonne dynamique, et fait montre d'un comportement très similaire à celui de $JI$. En fonction du paramétrage utilisé,

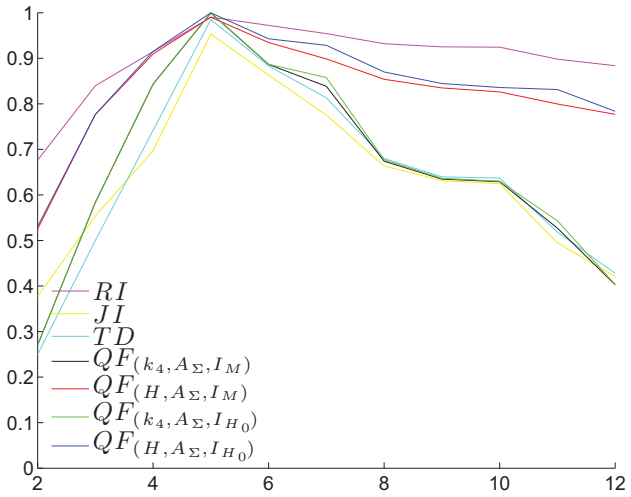



FIGURE 4.21 – Comparaison de partitions strictes $V_k$ à une partition stricte $U_{k^\star}$ de référence avec $RI$, $JI$, $TD$ et $QF$ en fonction de $k = 2, \ldots, 12$

notre propre indice $QF$ présente un comportement proche de celui de $JI$ et $TD$ ou de celui de $RI$. On constate que cette distinction repose sur la mesure de sparsité utilisée : la dynamique de $QF$ est meilleure lorsque celui-ci est calculé avec $\kappa_4$ car cette mesure fait montre d'un plus grand pouvoir discriminant, comme explicité et discuté à la page 110. Ensuite, on remarque que le choix de l'implication résiduelle n'a que peu d'influence sur $QF$ dans le cas de la comparaison de deux partitions strictes, mais nous verrons par les expérimentations suivantes que ce n'est pas toujours le cas. Enfin, une dernière remarque intéressante peut être faite à propos des faibles variations observées pour tous les indices entre certaines valeurs de $k$. Par exemple, et cela s'observe très bien pour les indices $JI$, $TD$, $QF(\kappa_4, \mathcal{A}_\sigma, I_M)$ et $QF(\kappa_4, \mathcal{A}_\sigma, I_{H_0})$, la concordance affichée par les indices pour la comparaison de $U_{k^\star}$ à $V_8$, $V_9$ et $V_{10}$ varie peu. Ces "plateaux" peuvent s'expliquer au regard de la Figure 4.2, page 82, où sont visualisés les clusters. On y voit que les partitions considérées partagent chacune, à de très rares points près, deux clusters en commun (orange et vert vif pour $V_8$, vert vif et magenta pour $V_9$, vert vif et bleu marine pour $V_{10}$) avec la partition de référence $U_{k^\star}$, si bien que malgré leur nombre de clusters croissant, leur accord avec $U_{c^\star}$ reste presque constant. Une analyse plus fine permettrait d'expliquer de la même manière les comportements certes moins prononcés, mais similaires, que l'on observe pour toutes les courbes de la Figure 4.21.

#### 4.4.3.2.b Données synthétiques – Partitions non strictes (E3)

Nous présentons et discutons ici les résultats de l'expérimentation **E3**, décrite page 83, consacrée à la comparaison d'une partition stricte de référence avec une

collection de partitions floues fournies par l'algorithme FCM et une collection partitions possibilistes fournies par l'algorithme PCM. Pour les indices considérés, ils sont donnés à la Figure 4.22.

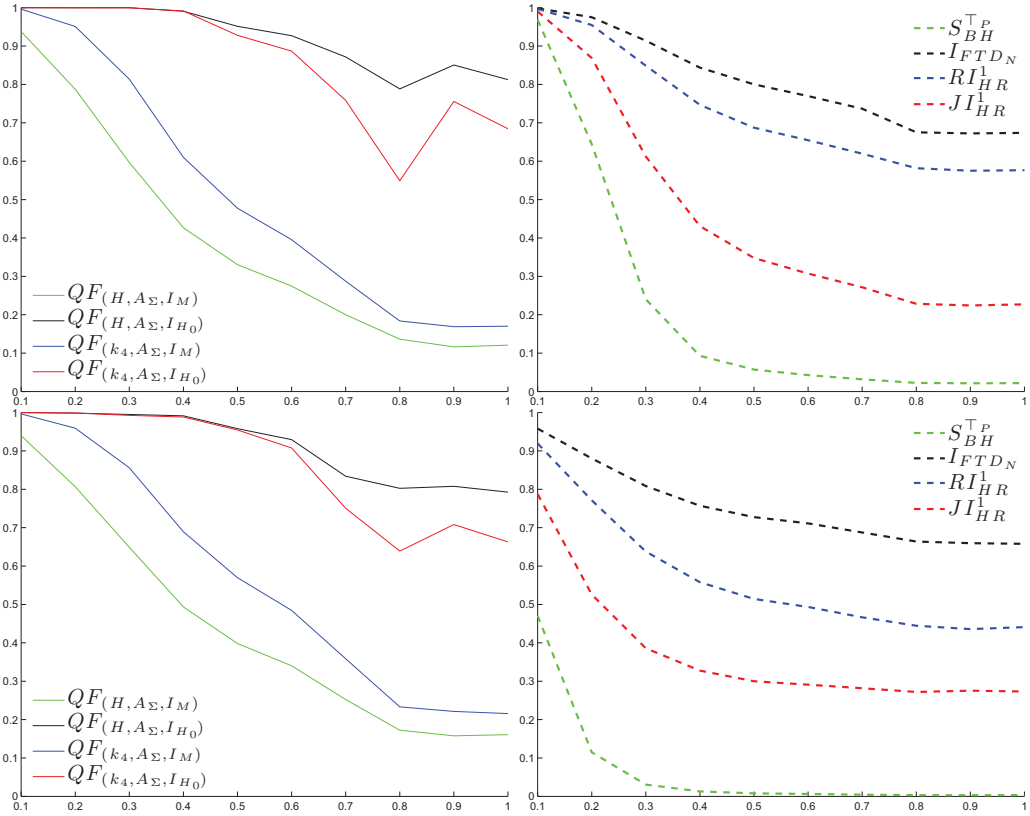



FIGURE 4.22 – Comparaison de deux collections de partitions non strictes à une partition stricte de référence $U_{k^\star}$. Première ligne : partitions floues $V_\sigma$. Deuxième ligne : partitions possibilistes $V'_\sigma$.

Notre indice se montre moins sensible à la nature des partitions que les indices de référence dont les valeurs sont plus faibles lorsque l'on considère des partitions possibilistes. On remarque que pour cette expérimentation, notre indice $QF_{(S,\mathcal{A},\mathcal{I}_\top)}$ est très sensible au choix de l'implication résiduelle floue, sa dynamique étant bien meilleure avec l'implication de Gödel $\mathcal{I}_M$ qu'avec celle de Hamacher $\mathcal{I}_{H_0}$, ce qui s'explique aisément. À mesure que $\sigma$ augmente, le chevauchement entre les clusters aussi, et les partitions $V_\sigma$ et $V'_\sigma$ sont ainsi de plus en plus floues. La matrice de contingence croisant chacune de ces partitions non strictes avec la partition stricte de référence est alors de moins en moins sparse, si bien que les valeurs fournies en entrée de l'implication résiduelle floue sont faibles. À deux valeurs faibles, $\mathcal{I}_M$ répond par une

valeur faible, tandis que $\mathcal{I}_{H_0}$ fonde sa valeur de sortie sur le rapport de ces deux valeurs. Si elles sont proches, l'implication aura une valeur élevée, de même l'indice. Les différences constatées entre les mesures de Hoyer $H$ et le kurtosis $\kappa_4$, s'opposent à celles constatées à l'expérimentation **E1**. Ceci peut s'expliquer par la propriété *(P1)* que satisfait $H$ mais pas $\kappa_4$. Ici, le nombre de clusters ne varie pas, tant et si bien qu'à mesure que $\sigma$ augmente, les cœfficients de la matrice de contigence croisant chaque partition $V_\sigma$ avec la partition de référence s'homogénéisent, ce qui revient à retirer un peu de l'amplitude des cœfficients les plus forts au profit des autres. Selon l'implication associée, et conformément aux remarques sur leur comportement respectif faites ci-dessus, il est logique qu'avec $H$, les courbes de l'indice soient les deux extrêmes.

#### 4.4.3.2.c Partitions non-strictes de données réelles (E4)

Nous nous attachons ici à discuter des résultats obtenus avec notre indice $QF_{(\mathcal{S},\mathcal{A},\mathcal{I}_\top)}$ et les confronter à ceux des indices de références $I_{FTD_N}$, $S^{\top_P}_{BH}$, $RI^1_{HR}$ et $JI^1_{HR}$ pour l'expérimentation **E4**, décrite page 83, dédiée à la comparaison de partitions non-strictes obtenues par clustering de jeux de données réelles. Les résultats obtenus avec notre indice sont donnés à la Figure 4.23 à la fois pour les partitions floues et partitions possibilistes. Une fois n'est pas coutume, nous renverrons le lecteur aux Figures 4.15 et 4.16 (pages 106 et 107) où les valeurs des indices de référence ont déjà été données.

Tout d'abord, on peut encore remarquer que notre indice est moins sensible à la nature des partitions comparées que ne le sont les indices de référence. Contrairement à ces derniers qui prennent une valeur maximale proche de 1 pour $c = c^\star$ car dans ce cas les partitions $U_{c^\star}$ (respectivement $U'_{c^\star}$) fournies par FCM (respectivement PCM) sont probablement proches de $R_{c^\star}$ (respectivement $R'_{c^\star}$), et bien qu'étant réflexif lui aussi, les paramétrages de notre indice avec l'implication résiduelle de Gödel $\mathcal{I}_M$ lui font réaliser son maximum à une valeur bien moindre, en comparaison à celle de $\mathcal{I}_{H_0}$. Ceci s'explique par le fait que la sortie de $\mathcal{I}_{H_0}$ est toujours inférieure ou égale à celle de $\mathcal{I}_{H_0}$. Par ailleurs, la dynamique est cette fois-ci meilleure avec l'implication de Hamacher $\mathcal{I}_{H_0}$, à mesure de sparsité donnée. Ceci s'explique simplement par la non linéarité de $\mathcal{I}_{H_0}$ combinée à la variation du nombre de clusters qui, en augmentant fait décroître la sparsité. Enfin, on remarque encore le pouvoir discriminant du Kurtosis $\kappa_4$ supérieur à celui de la mesure de Hoyer $H$ en termes de dynamique. Eût égard aux effets combinés des paramètres pour les situations spécifiques de chaque expérimentations, l'avantage revient à l'indice $QF_{(\kappa_4,\mathcal{A}_\Sigma,\mathcal{I}_{\top_M})}$ qui tire toujours son épingle du jeu.

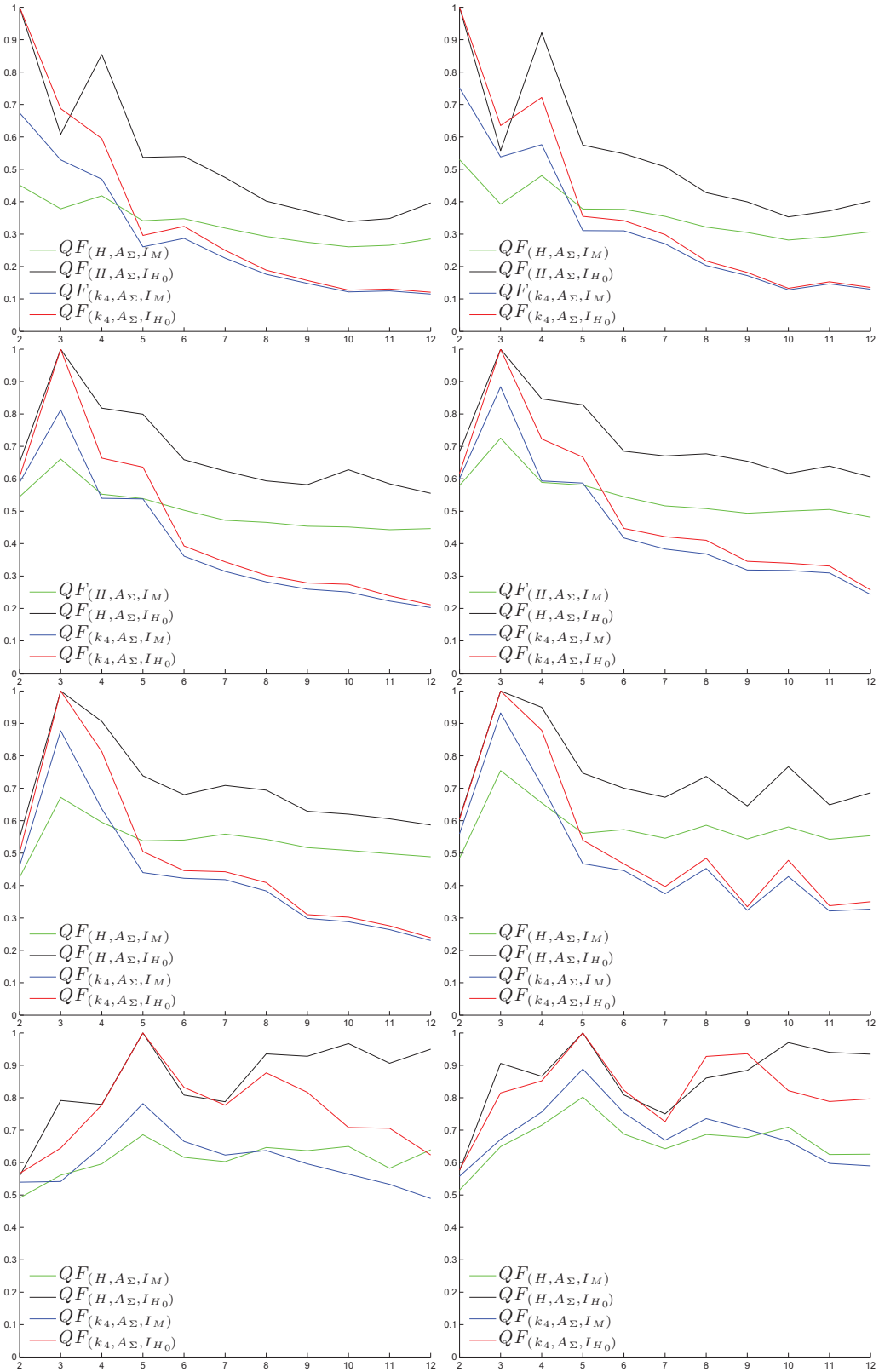



FIGURE 4.23 – Valeurs de $QF_{(\mathcal{S},\mathcal{A},\mathcal{F}_\top)}$ avec les différents paramétrages pour la comparaison de partitions non strictes avec une partition non stricte de référence. À gauche : partitions floues $U_c$ avec $R_{c^\star}$. À droite : partitions possibilistes $U'_c$ avec $R'_{c^\star}$, pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

## 4.5 Vers une présentation unifiée

Nous terminons ce chapitre par quelques réflexions menées pour l'élaboration d'un cadre unifiant toutes les approches de comparaison de deux partitions non-strictes orientées individus décrites dans ce mémoire, qu'elles soient natives ou fondées sur l'extension d'approches strictes. Les propositions de Campello [2007], Borgelt [2006] et de Hüllermeier et al. [2012] présentées respectivement aux pages 53, 60 et 65, mais aussi deux de nos propres propositions Quéré et al. [2010] et Quéré et Frélicot [2011a] présentées aux pages 85 et 96 du présent chapitre sont reformulées selon ce cadre. Ce dernier repose sur la présentation de chaque approche sous la forme d'un quadruplet de fonctions $\{f, f', g, N\}$ et complète celui que nous avons proposé dans [Quéré et Frélicot, 2011b], qui n'intégrait ni la proposition de Campello [2007], ni celle de Hüllermeier et al. [2012] et pour laquelle les approches étaient alors présentées sous la forme d'un triplet $\{f, g, N\}$. Toutefois, nous montrons que dans la plupart des cas $\{f, f', g, N\}$ se réduit à $\{f, g, N\}$.

### 4.5.1 Un cadre général pour la définition d'indices orientés individus

D'une manière ou d'une autre, toutes les approches de comparaison de deux partitions non-strictes orientées individus, qu'elles soient natives ou fondées sur l'extension d'approches strictes, reposent sur une même construction en deux étapes de calcul :

1) pour chaque paire d'individus, de degrés de coïncidence et/ou de non coïncidence dans une partition, dans l'autre,
2) d'un indice ou d'une distance duale qui agrège des termes d'accord/désaccord entre les deux partitions définis à partir de ces degrés.

#### 4.5.1.1 Étape 1 : le couple $\{f, f'\}$

La première étape consiste en la construction, pour chaque partition $U$ et $V$ à comparer, de deux matrices $\Psi$ et $\overline{\Psi}$, appelées respectivement *matrice de coïncidence* et *matrice de rejet*. Pour $U$, leur terme général respectif $\psi_{U,kl}$ et $\overline{\psi}_{U,kl}$ représentent le degré selon lequel les deux individus $\mathbf{x}_k$ et $\mathbf{x}_l$ appartiennent, respectivement n'appartiennent pas, au même cluster de $U$, et sont définis par :

$$\psi_{U,kl} = f(\mathbf{u}_k, \mathbf{u}_l) \quad \text{et} \quad \overline{\psi}_{U,kl} = f'(\mathbf{u}_k, \mathbf{u}_l), \tag{4.15}$$

où $f$ et $f'$ sont un couple $\{f, f'\}$ de fonctions : $[0,1]^c \times [0,1]^c \to [0,1]$, propre à chaque approche. La Table 4.7 recense les couples $\{f, f'\}$ que nous avons identifiés pour les propositions considérées. On remarque que les fonctions $f'$ utilisées par Borgelt, Hullermeier et al. et par Quéré et al. sont les complémentaires de leur fonction $f$, si

bien que le couple $\{f, f'\}$ peut être réduit à la seule fonction $f$, et le quadruplet réduit à un triplet $\{f, g, N\}$ comme formulé dans [Quéré et Frélicot, 2011b].

TABLE 4.7 – Fonctions $f$ et $f'$ pour différentes approches

| Approche | $f$ | $f'$ |
|---|---|---|
| Campello [2007] | $\bigperp_{i=1}^{c} \top(a_i, b_i)$ | $\bigperp_{\substack{i \neq j \\ i,j=1}}^{c} \top(a_i, b_j)$ |
| Borgelt [2006] | $\sum_{i=1}^{c} \top(a_i, b_i)$ | $1 - f(\mathbf{a}, \mathbf{b})$ |
| Quéré et al. [2010] et Quéré et Frélicot [2011a] | $\frac{\sum_{i=1}^{c} \top(a_i, b_i)}{\top\left(K_\top(\sum_{i=1}^{c} \top(a_i, a_i)), K_\top(\sum_{i=1}^{c} \top(b_i, b_i)\right)}$ | $1 - f(\mathbf{a}, \mathbf{b})$ |
| Hüllermeier et al. [2012] | $1 - d(\mathbf{a}, \mathbf{b})$ | $1 - f(\mathbf{a}, \mathbf{b})$ |

#### 4.5.1.2 Étape 2 : le couple $\{g, N\}$

La seconde et dernière étape de la construction consiste en la définition d'un indice de comparaison non strict $I$ par :

$$I(U, V) = \frac{1}{N} \sum_{k=2}^{n} \sum_{l=1}^{k-1} g_{m_{\alpha\beta}}(k, l), \tag{4.16}$$

où $N \in \mathbb{R}_+$ est un facteur de normalisation et où $g$ est une fonction linéaire de la contribution aux quatre termes $\{m_{\alpha\beta} : \alpha, \beta \in \{0, 1\}\}$ d'accord/désaccord de la matrice de contingence-paires $M(U, V)$, de la paire d'individus $\{\mathbf{x}_k, \mathbf{x}_l\}$, calculés à partir des matrices $\Psi$ et $\overline{\Psi}$ de chacune des deux partitions :

$$m_{11}(k, l) = f_{11}(\psi_{U,kl}, \psi_{V,kl}), \tag{4.17}$$
$$m_{10}(k, l) = f_{10}(\psi_{U,kl}, \overline{\psi}_{V,kl}), \tag{4.18}$$
$$m_{01}(k, l) = f_{01}(\overline{\psi}_{U,kl}, \psi_{V,kl}), \tag{4.19}$$
$$m_{00}(k, l) = f_{00}(\overline{\psi}_{U,kl}, \overline{\psi}_{V,kl}), \tag{4.20}$$

où $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ est un quadruplet de fonctions : $[0, 1] \times [0, 1] \to [0, 1]$. Pour toutes les propositions considérées, les termes de désaccord sont symétriques, de sorte que $f_{10} = f_{01}$. Les quadruplets sont donnés à la Table 4.9. De même, la Table 4.8 fournit quelques exemples de couples $\{g_{m_{\alpha\beta}}, N\}$ permettant de retrouver certains indices dérivés de ces approches. On remarquera que leur expression est à peu de choses près la même que celle des indices stricts fondés sur une approche orientée individus, présentés au Chapitre 2 à la Table 2.3, page 26, si bien qu'il est aisé de trouver les couples

$\{g_{m_{\alpha\beta}}, N\}$ permettant de dériver tous les indices stricts de cette famille. Dans [Quéré et Frélicot, 2011b], nous nous concentrions sur l'expression directe des indices plutôt que sur la formulation générale permettant de dériver toute mesure d'appariement, si bien que les fonctions $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ n'étaient pas explicitées. Toutefois, les couples $\{g, N\}$ présentés dans cet article sont strictement les mêmes que ceux donnés ici.

### 4.5.2 À propos des propriétés de $\{f, f'\}$, $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ et $\{g, N\}$

#### 4.5.2.1 Étape 1 : propriétés du couple $\{f, f'\}$

Pour générer une matrice de coïncidence d'une partition $U \in \mathbb{M}_{pcn}$, les fonctions $f$ et $f'$ doivent nécessairement satisfaire les conditions aux bornes suivantes : $\forall \mathbf{a}, \mathbf{b} \in \{0,1\}^c$,

$$f(\mathbf{a}, \mathbf{b}) = 1 \text{ et } f'(\mathbf{a}, \mathbf{b}) = 0 \qquad \text{si } a_i = b_i \ \forall i = 1, c \tag{4.21}$$

$$f(\mathbf{a}, \mathbf{b}) = 0 \text{ et } f'(\mathbf{a}, \mathbf{b}) = 1 \qquad \text{si } a_i \neq b_i \ \forall i = 1, c. \tag{4.22}$$

Les fonctions $f$ et $f'$ doivent aussi être symétriques :

$$f(\mathbf{a}, \mathbf{b}) = f(\mathbf{b}, \mathbf{a}) \text{ et } f'(\mathbf{a}, \mathbf{b}) = f'(\mathbf{b}, \mathbf{a}). \tag{4.23}$$

Tout couple de fonctions $\{f, f'\}$ satisfaisant ces conditions permettent de retrouver respectivement la matrice de coïncidence stricte $\Psi_U$ telle que définie par (2.27), page 29, et la matrice $(\mathbb{1}_{nn} - \Psi_U)$ complémentaire à cette dernière lorsque $U$ est une partition stricte de $\mathbb{M}_{hcn}$. Il peut aussi être souhaitable, mais non nécessaire, que $f$ et $f'$ soient complémentaires :

$$f(\mathbf{a}, \mathbf{b}) + f'(\mathbf{a}, \mathbf{b}) = 1, \tag{4.24}$$

si bien qu'alors, le degré $f(\mathbf{u_k}, \mathbf{u_l}) + f'(\mathbf{u_k}, \mathbf{u_l})$ selon lequel les individus $\mathbf{x}_k$ et $\mathbf{x}_l$ appartiennent et n'appartiennent pas au même cluster de $U$ soit total, ce qui semble conforme à l'intuition. On remarquera que tel n'est pas le cas pour la proposition de [Campello, 2007], comme le montre la Table 4.7.

#### 4.5.2.2 Étape 2 : propriétés du couple $\{g, N\}$ et des fonctions $\{f_{11}, f_{10}, f_{01}, f_{00}\}$

Les fonctions $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ doivent satisfaire les conditions aux bornes suivantes :

$$f_{11}(1,1) = 1, \quad f_{11}(0,0) = 1, \quad f_{11}(1,0) = 0, \quad f_{11}(0,1) = 0, \tag{4.25}$$

$$f_{10}(0,1) = 1, \quad f_{10}(1,0) = 1, \quad f_{10}(0,0) = 0, \quad f_{10}(1,1) = 0, \tag{4.26}$$

TABLE 4.8 – Fonctions $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ pour différentes approches

| Approche | $f_{11}$ | $f_{10}$ | $f_{01}$ | $f_{00}$ |
|---|---|---|---|---|
| Campello [2007] | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ |
| Borgelt [2006] | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ |
| Quéré et al. [2010] | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ | $\top(a,b)$ |
| Quéré et Frélicot [2011a] | $s(a,b) \times \mathcal{A}(a,b)$ | $(1-s(a,1-b)) \times \mathcal{A}(a,b)$ | $(1-s(a,1-b)) \times \mathcal{A}(a,b)$ | $s(a,b) \times \mathcal{A}(a,b)$ |
| Hüllermeier et al. [2012] | $s(a,b) \times ab$ | $\top_{\text{Ł}}(a,b)$ | $\top_{\text{Ł}}(a,b)$ | $s(a,b) \times (a+b-ab)$ |

avec $s(a,b) = 1 - \mid a - b \mid$ et $\mathcal{A}(a,b)$ la moyenne arithmérique.

TABLE 4.9 – Couples $\{g_{m_{\alpha\beta}}, N\}$ permettant de dériver quelques indices

| Indice | $g_{m_{\alpha\beta}}$ | $N$ |
|---|---|---|
| *RI* | $m_{11}(k,l) + m_{00}(k,l)$ | $\sum_{k=2}^{n} \sum_{l=1}^{k-1} \left( m_{11}(k,l) + m_{10}(k,l) + m_{01}(k,l) + m_{00}(k,l) \right)$ |
| *JI* | $m_{11}(k,l)$ | $\sum_{k=2}^{n} \sum_{l=1}^{k-1} \left( m_{11}(k,l) + m_{10}(k,l) + m_{01}(k,l) \right)$ |
| *RAI* | $m_{11}(k,l) + m_{10}(k,l) + m_{00}(k,l)$ | $\sum_{k=2}^{n} \sum_{l=1}^{k-1} \left( m_{11}(k,l) + m_{10}(k,l) + m_{01}(k,l) + m_{00}(k,l) \right)$ |

$$f_{01}(0,1) = 1, \qquad f_{01}(1,0) = 1, \qquad f_{01}(0,0) = 0, \qquad f_{01}(1,1) = 0, \tag{4.27}$$

$$f_{00}(1,1) = 1, \qquad f_{00}(0,0) = 1, \qquad f_{00}(1,0) = 0, \qquad f_{00}(0,1) = 0. \tag{4.28}$$

Ces conditions permettent d'assurer que les indices dérivés se ramènent à leur pendant strict lors de la comparaison de deux partitions de $\mathbb{M}_{hcn}$. En effet, si tel est le cas, on a :

$$\sum_{k=2}^{n}\sum_{l=1}^{k-1} m_{\alpha\beta}(k,l) = m_{\alpha\beta}, \tag{4.29}$$

où $\{m_{\alpha\beta} : \alpha,\beta \in \{0,1\}\}$ sont les termes de la matrice de contingences-paire $M(U,V)$ croisant deux partitions strictes, définie page 18. Par ailleurs, il peut être intéressant que ces fonctions soient telles que, pour toutes les partitions $U$ et $V$ de $\mathbb{M}_{pcn}$ :

$$\sum_{k=2}^{n}\sum_{l=1}^{k-1} m_{11}(k,l) + m_{10}(k,l) + m_{01}(k,l) + m_{00}(k,l) = \frac{n(n-1)}{2} = q. \tag{4.30}$$

Lorsque $U$ et $V$ sont strictes, cette propriété est toujours satisfaite grâce aux conditions aux bornes (4.25), (4.26), (4.27) et (4.28).

Les fonctions $f_{11}$ et $f_{00}$ peuvent aussi être duales :

$$f_{11}(a,a) + f_{00}(1-a,1-a) = 1. \tag{4.31}$$

Comme nous l'avons évoqué plus tôt, la fonction $g$ et le facteur de normalisation $N$ forment quant à eux un couple $\{g,N\}$ dont l'expression dépend de l'indice $I$ que l'on veut dériver. Concernant $N$, il doit être choisi de sorte que :

(*N-1*) $I$ défini par (4.16) soit à valeurs dans $[0,1]$,

(*N-2*) $I(U,U) = 1$ si $U \in \mathbb{M}_{hcn}$.

L'intérêt de ce cadre unifié est d'étudier les propriétés satisfaites par tout indice $I$ dérivé selon plusieurs propriétés supplémentaires satisfaites par les fonctions $\{f,f'\}$ ainsi que par les éléments du couple $\{g,N\}$.

**Définition 27.** *[Zadeh, 1971] Une relation d'equivalence floue sur un ensemble $X$ est une fonction $e : X \times X \to [0,1]$ telle que, pour tout $\mathbf{a},\mathbf{b},\mathbf{c} \in X$ :*

(E-1) $e(\mathbf{a},\mathbf{a}) = 1$ *(réflexivité),*

(E-2) $e(\mathbf{a},\mathbf{b}) = e(\mathbf{b},\mathbf{a})$ *(symétrie),*

(E-3) $e(\mathbf{a},\mathbf{b}) \geq \top\big(e(\mathbf{a},\mathbf{c}), e(\mathbf{c},\mathbf{b})\big)$ *(⊤-transitivité).*

**Théorème 7.** *[De Baets et Mesiar, 1997] Si $f$ est une relation d'équivalence floue $\top_{Ł}$-transitive (Définition 27), et si $f$ et $f'$ se complémentent à 1 (4.24), alors $f'$ est une pseudométrique.*

**Théorème 8.** *Si $f$ et $f'$ se complémentent à 1 (4.24), et si $f_{11}$ et $f_{00}$ sont duales (4.31), alors $g_{m_{\alpha\beta}}(k,l) = m_{11}(k,l) + m_{00}(k,l)$ est réflexive (E-1).*
*De plus, si $N$ vérifie (N-2), alors l'indice $I$ construit à partir de $\{f, g_{m_{\alpha\beta}}, N\}$ est lui-même réflexif (II-1).*

**Théorème 9.** *La distance duale $D_I$ de l'indice $I$ construit à partir du triplet $\{f, g_{m_{\alpha\beta}}, N = q\}$ est une pseudométrique si et seulement si :*

- *$f$ et $f'$ sont conformes à la Proposition 7,*
- *$g_{m_{\alpha\beta}}(k,l) = m_{11}(k,l) + m_{00}(k,l)$ est réflexive (E-1) et $\top_{Ł}$ – transitive (E-3).*

À titre d'illustration, nous avons reporté les propriétés satisfaites par les indices de Rand dérivés de chacune des approches considérées à la Table 4.10. On y voit en particulier que le dual de l'indice $RI^1_{HR}$ de Hüllermeier et al. [2012] est ainsi une pseudométrique, puisque satisfaisant l'ensemble des propriétés requises. Notre indice $RI^{\top}_{NSW}$, proposé dans [Quéré et Frélicot, 2011a], constitue quant à lui une famille d'indices dont les duals sont toujours des pseudométriques lorsque $f$ est calculée avec la t-norme produit $\top_P$ et lorsque le paramètre de courbure $r$ du profil (4.11) de la fonction de similarité (4.10) est choisi nul (voir page 97).

## 4.6 Conclusion

Dans ce chapitre, nous avons présenté nos contributions au domaine de la comparaison de partitions non-strictes. Trois nouvelles constructions ont été décrites et comparées aux propositions majeures de la littérature revues au chapitre précédent. La première de ces constructions [Quéré et al., 2010] est fondée sur la généralisation à toute norme triangulaire de la proposition de Brouwer visant à corriger le comportement contre-intuitif de la proposition de Borgelt. Il s'agit donc d'une extension, orientée individus. Notre deuxième construction, native et orientée individus elle aussi, consiste en la définition d'une nouvelle fonction de similarité pour l'agrégation des matrices de coïncidence, menant à la définition d'un nouvel indice de comparaison [Quéré et Frélicot, 2011a]. Enfin, la troisième et dernière construction est une approche native orientée clusters, qui repose sur une mesure de la sparsité de la matrice de contingence pour la définition d'un nouvel indice [Quéré et Frélicot, 2012].

TABLE 4.10 – Propriétés satisfaites par l'indice de Rand dérivé de différentes approches, pour deux partitions de $\mathbb{M}_{fcn}$

| Propriété | | $RI_C^{\top}$ | $RI_B^{\top}$ | $RI_Q^{\top}$ | $RI_{HR}^{1}$ | $RI_{NSW}^{\top}$ |
|---|---|---|---|---|---|---|
| $f$ | dualité | | ● | ● | ● | ● |
| | réflexivité | | ○ | ● | ● | ● |
| | $\top_Ł$-transitivité | | ○ | ○ | ● | ○ |
| $f_{11}, f_{10}, f_{01}, f_{00}$ | propriété (4.30) | ● | ● | ● | ● | ● |
| $g$ | réflexivité | | | | ● | ● |
| | symétrie | ● | ● | ● | ● | ● |
| | $\top_Ł$-transitivité | | | | ● | ◇ |
| $N$ | $N = q$ | | | | ● | ● |

● : propriété satisfaite,
● : propriété satifaite uniquement avec $\top_P$,
○ : propriété satisfaite uniquement avec $\top_M$,
◇ : propriété satisfaite uniquement avec le paramètre $r = 0$.

Pour chacune de ces propositions, les résultats obtenus à partir d'une même série d'expérimentations ont été présentés et analysés. Ces expérimentations ont été choisies pour couvrir au mieux l'ensemble des aspects de la problématique et étudier le comportement des indices considérés lorsque soumis à différentes situations. Ainsi, nous nous sommes attachés à observer et décrire la sensibilité de chaque mesure au degré de flou ainsi qu'au nombre de clusters des partitions comparées. Nous avons donné les complexités afin de situer nos propositions par rapport à celles de la littérature (Table 4.11). Nous avons aussi étudié l'incidence de la nature des partitions comparées, alimentant ainsi la discussion entamée au chapitre précédent au sujet de la comparaison de partitions issues de différents espaces (Table 4.12). Les résultats présentés ont permis de montrer que nos propositions présentent des comportements intéressants et surpassent souvent les approches existantes de la littérature de la même catégorie auxquelles elles sont comparées. Nous avons montré la grande flexibilité de nos propositions permise par leur paramétrages respectifs qui offre de nombreuses possibilités au praticien et notamment celle de calquer le comportement de plusieurs mesures de la littérature aux spécificités propres[8].

8. Par exemple, notre proposition fondée sur une fonction de similarité peut adopter un comportement proche des indices de Rand et de Jaccard.

Enfin, nous avons dessiné les prémisses d'un cadre unifiant la construction de toutes les approches orientées individus, natives ou non, en identifiant les diverses propriétés que doivent ou peuvent satisfaire ces approches, dans la suite des travaux de Quéré et Frélicot [2011a]. Ces propriétés ont été discutées et partiellement reliées aux propriétés métriques souhaitées pour tout indice de comparaison. Afin d'illustrer nos propos, les propositions concernées décrites dans ce mémoire ont été exprimées selon ce formalisme. Bien qu'encore incomplet, un tel cadre nous semble prometteur car il ouvre la voie à la définition d'indices plus performants et dont les propriétés pourraient être connues et choisies ad hoc, contrairement à nombre de propositions de la littérature.

TABLE 4.11 – Complexités en temps et en espace des mesures revues en fonction de deux partitions de taille $c \times n$ et $r \times n$.

| Mesure de comparaison | Complexité en temps | Complexité en espace |
|---|---|---|
| Campello (page 53) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Anderson (page 56) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Ceccarelli (page 58) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Borgelt (page 60) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Brouwer (page 63) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| FTD (page 64) | $\mathcal{O}(n)$ | $\mathcal{O}(max(c,r)^2)$ |
| Hüllermeier et al. (page 65) | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Beringer (page 68) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Runkler (page 70) | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |
| Bodjanova (page 70) | $\mathcal{O}(n)$ | $\mathcal{O}(c^2)$ |
| Acciani (page 71) | $\mathcal{O}(n)$ | $\mathcal{O}(max(c,r)^2)$ |
| DNC (page 73) | $\mathcal{O}(n)$ | $\mathcal{O}(n)$ |
| Quéré et al. [2010] | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Quéré et Frélicot [2011a] | $\mathcal{O}(n^2)$ | $\mathcal{O}(n^2)$ |
| Quéré et Frélicot [2012] | $\mathcal{O}(n)$ | $\mathcal{O}(cr)$ |

TABLE 4.12 – Espaces de définition des mesures revues

| Mesure de comparaison | $\mathbb{M}_{hcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}_{fcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}^{\leq}_{pcn}$ | $\mathbb{M}_{hcn} \times \mathbb{M}^{>}_{pcn}$ | $\mathbb{M}_{fcn}$ | $\mathbb{M}^{\leq}_{pcn}$ | $\mathbb{M}^{>}_{pcn}$ |
|---|---|---|---|---|---|---|---|
| Campello (page 53) | ⊗ | ⊗ | ⊗ | ⊗ | × | × | × |
| Anderson (page 56) | ⊖ | ⊖ | ⊖ | ⊖ | ⊖ | ⊖ | ⊖ |
| Ceccarelli (page 58) | × | × | × | × | ⊗ | × | × |
| Borgelt (page 60) | ⊗ | ⊗ | × | | ⊗ | × | |
| Brouwer (page 63) | ⊗ | ⊗ | × | | ⊗ | × | |
| FTD (page 64) | × | × | × | × | ⊗ | × | × |
| Hüllermeier et al. (page 65) | ⊗ | ⊗ | × | | ⊗ | × | |
| Beringer (page 68) | × | × | × | × | ⊗ | × | × |
| Runkler (page 70) | × | × | × | × | ⊗ | × | × |
| Bodjanova (page 70) | × | × | × | × | ⊗ | × | × |
| Acciani (page 71) | × | × | × | × | ⊗ | × | × |
| DNC (page 73) | × | ⊗ | × | × | | | |
| Quéré et al. [2010] | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | |
| Quéré et Frélicot [2011a] | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | |
| Quéré et Frélicot [2012] | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ |

⊗ : La mesure est définie par ses auteurs pour une telle comparaison.

⊖ : La mesure est définie par ses auteurs pour une telle comparaison mais présente un défaut de construction qui empêche de l'utiliser l'état.

× : La mesure peut techniquement être utilisée pour une telle comparaison.

# CHAPITRE 5 :

# Conclusion et perspectives

Dans ce chapitre terminal, nous résumons nos travaux sur la comparaison de partitions non-strictes, rappelons nos principales propositions, puis nous dressons quelques perspectives pour de futurs travaux.

**Apports scientifiques**

À travers les chapitres 2 et 3, respectivement consacrés à l'état de l'art aussi complet que possible des mesures de comparaison de partition strictes et non-strictes, nous avons proposé une taxonomie différenciant les approches *orientées individus* et celles *orientées clusters*, qui permet de rapprocher la plupart des propositions de mesure existantes d'un point de vue technique mais aussi théorique, selon la notion de concordance entre partitions considérée. Les propriétés métriques de chaque mesure de comparaison ont été systématiquement qualifiées, même lorsque les auteurs n'en ont pas discuté. Dans le chapitre 3, nous avons par ailleurs décrit des liens existant entre certaines mesures, liens absents de la littérature. Enfin, dans ce même chapitre, nous avons étudié chacune des approches non-strictes revues au regard des différents espaces de partitions (strict, flou et possibiliste) pour lesquelles elles sont définies, et nous avons discuté du problème ouvert de la comparaison de partitions issues d'espaces différents.

Dans le chapitre 4, nous avons proposé trois mesures de comparaison de partitions non strictes. La première est une *extension de mesures strictes orientée individus* [Quéré et al., 2010]. Elle généralise à toute t-norme une proposition existante vouée à corriger le comportement contre-intuitif d'une approche majeure de la littérature. La deuxième, qui repose en partie sur la première, est une *mesure native orientée individus* [Quéré et Frélicot, 2011a] fondée sur une mesure de similarité entre les coïncidences de chacune des deux partitions à comparer. Cette mesure répond au problème de la non-réflexivité observée pour un grand nombre de mesures non-

strictes de la littérature, et contrairement à ces dernières, elle prend donc sa valeur maximale lors de la comparaison d'une partition floue avec elle-même. De plus, elle fait montre d'une grande flexibilité et permet ainsi au praticien d'en ajuster le comportement en fonction de la situation et de ses besoins. Enfin, la troisième mesure que nous avons proposée est une mesure *native orientée clusters* reposant sur la mesure de la sparsité des matrices de contingence croisant deux partitions non strictes [Quéré et Frélicot, 2012] . Elle est aussi réflexive, et présente l'avantage de présenter une très faible complexité en temps, en faisant une mesure privilégiée pour la comparaison de partitions de grands ensembles d'individus. Chacune de ces mesures a fait l'objet d'une étude expérimentale dont les résultats ont montré leur compétitivité au regard des autres mesures de la littérature, surpassant plusieurs d'entre elles. À la fin de ce chapitre, nous avons esquissé les prémisses d'un formalisme pour la définition des mesures non-strictes orientées individus au travers duquel les propriétés des mesures s'y conformant sont plus aisément identifiables, facilitant ainsi leur comparaison. À titre d'exemple, nous avons exprimé et comparé plusieurs des mesures revues dans ce mémoire selon ce cadre.

**Pistes de travail**

Il ressort des travaux de recherche présentés dans ce mémoire plusieurs perspectives.

Tout d'abord, même si la concordance et la discordance totales entre deux partitions sont bien cernées, la concordance partielle reste une notion assez vague. Tant et si bien que l'on doit souvent s'en remettre à l'intuition pour statuer sur la concordance exhibée. Préciser cette notion est une perspective exaltante. Pour les partitions floues, une piste à explorer serait la caractérisation des relations binaires floues inhérentes à toute partition de ce type, et qui intervient notamment dans la construction des approches orientées individus par l'intermédiaire de la matrice de coïncidence. Il pourrait être ainsi intéressant de relier ces approches avec les travaux de Ovchinnikov [1991], dans lesquels une partition floue est formellement définie à partir des classes d'équivalence d'une relation d'équivalence floue. En réalisant un tel lien, nous pensons pouvoir préciser la notion d'équivalence entre deux partitions floues et établir un cadre formel pour la définition de mesures de comparaisons orientées individus, complétant celui que nous avons esquissé dans [Quéré et Frélicot, 2011b].

Toujours selon cette idée, nous avons de sérieuses raisons de penser que la perte de la propriété d'identité des indiscernables, observée pour toutes les approches orientées individus, peut s'expliquer par l'existence d'une équivalence ainsi définie entre les partitions comparées. Cette perte est induite par le fait qu'il est possible de trouver

deux partitions différentes présentant la même matrice de coïncidence, si bien que selon ces approches, ces partitions sont totalement concordantes. Or, nous avons observé numériquement que de telles partitions sont liées par une matrice bistochastique orthogonale, dont la matrice de permutation permettant l'appariement des clusters de chacune des partitions n'est qu'un cas particulier. Nous pensons que ces matrices décrivent une transformation géométrique préservant les structures de groupe propres aux partitions. Si tel est le cas, il serait alors naturel de penser que deux partitions ainsi liées sont strictement équivalentes, puisque décrivant les mêmes relations entre les individus qu'elles caractérisent.

Ensuite, la taxonomie adoptée dans ce mémoire met à jour de fortes similitudes entre plusieurs propositions de mesures orientées clusters. Nombre d'entre elles reposent en particulier sur l'agrégation d'une mesure de compatibilité entre les clusters de chaque partition, si bien que nous avons commencé un travail sur la définition d'un cadre unifiant toutes ces mesures qui devrait voir le jour à très court terme.

Enfin, en mettant de côté l'aspect sémantique qui nous l'a interdit, le problème de la comparaison d'une partition floue avec une partition possibliste reste un problème mathématique ouvert. Une piste consiste à transformer la partition possibiliste en une partition floue que l'on pourrait alors comparer à l'aide d'une mesure définie pour les partitions floues. La transformation peut s'avérer très différente selon que la partition est dans $\mathbb{M}_{pcn}^{>}$ ou dans $\mathbb{M}_{pcn}^{\leq}$, et peut constituer un problème inverse.

## Références de l'auteur


Quéré, R. et C. Frélicot. A new index based on sparsity measures for comparing fuzzy partitions. To appear in 9th Int. Workshop in Statistical Techniques in Pattern Recognition, SPR. Hiroshima, Japan, 2012.

Quéré, R. et C. Frélicot. Une extension de l'indice de Rand par une mesure de similarité entre matrices de partitions non strictes. In Rencontres de la Société Francophone de Classification, SFC. Orléans, France, 2011.

R. Quéré et C. Frélicot. A normalized soft window-based similarity measure to extend the Rand index. In 20th IEEE Int. Conf. on Fuzzy Systems, FUZZ-IEEE. Taipei, Taïwan, 2011.

R. Quéré et C. Frélicot. A general framework for a class of comparison indices of soft partitions. In 14th IFSA World Congress. Surabaya, Indonesia. 2011.

Quéré, R. et R. Pétéri. Suivi robuste de poissons dans des passes à poissons par filtrage particulaire et transformée en curvelets. 23rd GRETSI conference. Bordeaux, France, 2011.

R. Quéré, H. Le Capitaine, N. Fraisseix, et C. Frélicot. On normalizing fuzzy coincidence matrices to compare fuzzy and/or possibilistic partitions with the rand index. In 10th. IEEE Int. Conf. on Data Mining, ICDM. Sydney, Australia, 2010.

Quéré, R., H. Le Capitaine, N. Fraisseix et C. Frélicot (2010). Sur la normalisation des matrices de coïncidence pour comparer deux partitions floues. In Rencontres Francophones sur la Logique Floue et ses Applications, LFA. Lannion, France, 2010.

# Bibliographie


G. Acciani, G. Fornarelli, et L. Litturi. Comparing fuzzy data sets by means of graph matching technique. In *Lecture Notes in Computer Science 2714*, 2003. [Cité page 71]

A. N. Albatineh, M. Niewiadomska-Bugaj, et D. Mihalko. On similarity indices and correction for chance agreement. *J. of Classification*, 23, 2006. [Cité page 25]

D. T. Anderson, J. C. Bezdek, M. Popescu, et J. M. Keller. Comparing fuzzy, probabilistic, and possibilistic partitions. *IEEE Trans. on Fuzzy Systems*, 18(5), 2010. [7 citations pages 7, 29, 52, 56, 57, 89, et 114]

P. Arabie et S. A. Boorman. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10 :148–203, 1973. [6 citations pages 1, 12, 13, 16, 17, et 38]

M. Baczyński et B. Jayaram. (s, n)- and r-implications : A state-of-the-art survey. *Fuzzy Sets Syst.*, 159(14) :1836–1859, July 2008. ISSN 0165-0114. [Cité page 50]

J. Beringer et E. Hüllermeier. Fuzzy clustering of parallel data streams. In J. Valente de Oliveira et W. Pedrycz, editors, *Advances in Fuzzy Clustering and Its Application*, pages 333–352. John Wiley and Sons, 2007. [4 citations pages 68, 70, 101, et 114]

J. Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, 1981. [5 citations pages 7, 11, 80, 145, et 146]

J. C. Bezdek, R. J. Hathaway, et J. M. Huband. Visual assessment of clustering tendency for rectangular dissimilarity matrices. *Trans. Fuz Sys.*, 15(5) :890–903, October 2007. ISSN 1063-6706. [Cité page 12]

S. Bodjanova. Comparison of fuzzy partitions based on their $\alpha$-cuts. *Fuzzy Sets and Systems*, 105(1), 1999. [Cité page 70]

C. Borgelt. Finding the number of fuzzy clusters by resampling. In *16th IEEE Int. Conf. on Fuzzy Systems*, 2006. [8 citations pages 53, 60, 66, 67, 89, 121, 122, et 124]

B. Bouchon-Meunier, M. Rifqi, et S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84(2) :143 – 153, 1996. ISSN 0165-0114. [Cité page 52]

R. L. Brennan et R. J. Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27 :154–163, 1974. [2 citations pages 12 et 38]

R. K. Brouwer. Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *J. of Intelligent Information Systems*, 32(3), 2009a. [4 citations pages 28, 53, 63, et 85]

R. K. Brouwer. Permutation clustering using the proximity matrix. In *Proceedings of the 18th international conference on Fuzzy Systems*, FUZZ-IEEE'09, pages 441–446, Piscataway, NJ, USA, 2009b. IEEE Press. ISBN 978-1-4244-3596-8. [Cité page 12]

R. J. G. B. Campello. A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7), 2007. [10 citations pages 25, 52, 53, 60, 65, 89, 121, 122, 123, et 124]

R. J. G. B. Campello. Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters*, 31(9), 2010. [4 citations pages 62, 64, 101, et 114]

M. Ceccarelli et A. Maratea. A fuzzy extension of some classical concordance measures and an efficient algorithm for their computation. In *12th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems*, 2008. [4 citations pages 21, 53, 58, et 83]

I. Charon, L. Denoeud, A. Guenoche, et O. Hudry. Maximum transfer distance between partitions. *J. of Classification*, 23(1), 2006. [2 citations pages 34 et 64]

I. Charon, L. Denoeud, et O. Hudry. Maximum de la distance de transfert à une partition donnée. *Mathématiques & Sciences Humaines*, 179 :45–83, 2007. [2 citations pages 1 et 12]

M. Chavent, C. Lacomblez, et B. Patouille. Critère de rand asymétrique. In *Congrès de la SFC*, 2001. [6 citations pages 1, 12, 17, 23, 24, et 27]

R. N. Davé et R. Krishnapuram. Robust clustering methods : an unified view. *IEEE Trans. on Fuzzy Systems*, 5(2), 1997. [Cité page 11]

W. H. E. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, 1(3) :269–287, 1981. ISSN 0165-4896. [2 citations pages 34 et 36]

B. De Baets et R. Mesiar. Pseudo-metrics and t-equivalences. *J. Fuzzy Mathematics*, 5 :471–481, 1997. [Cité page 126]

L. Denoeud et A. Guenoche. Comparison of distance indices between partitions. In *Proceedings of IFCS'2006, Data Science and Classification*, 2006. [Cité page 34]

A. G. Di Nuovo et V. Catania. On external measures for validation of fuzzy partitions. In *Lecture Notes in Computer Science 4529*, 2007. [Cité page 73]

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, July 1945. [Cité page 26]

S. Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical report, Amsterdam, The Netherlands, The Netherlands, 2000. [Cité page 33]

J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J. of Cybernetics*, 3, 1973. [Cité page 11]

L. A. Fono, H. Gwet, et B. Bouchon-Meunier. Fuzzy implication operators for difference operations for fuzzy sets and cardinality-based measures of comparison. *European Journal of Operational Research*, 183(1) :314–326, 2007. [Cité page 51]

E. B. Fowlkes et C. L. Mallows. A method for comparing two hierarchical clusterings. *J. of the American Statistical Association*, 78, 1983. [4 citations pages 12, 22, 24, et 26]

A. Frank et A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`. [2 citations pages 79 et 83]

I. Gath et A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7) :773–781, July 1989. ISSN 0162-8828. [Cité page 8]

J. C. Gower et P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1) :5–48, 1986. ISSN 0176-4268. [Cité page 26]

M. Grabisch, J. Marichal, R. Mesiar, et E. Pap. *Aggregation Functions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009. [Cité page 113]

D. Gusfield. Partition-distance : A problem and class of perfect graphs arising in clustering. *Inf. Process. Lett.*, pages 159–164, 2002. [Cité page 65]

T. M. Ha. The optimum class-selective rejection rule. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(6) :608–615, June 1997. ISSN 0162-8828. [Cité page 73]

RR. J. Hathaway et J. C. Bezdek. Visual cluster validity for prototype generator clustering models. *Pattern Recognition Letters*, 24(9–10) :1563 – 1569, 2003. ISSN 0167-8655. [Cité page 12]

L. Hubert et P. Arabie. Comparing partitions. *J. of Classification*, 2(1), 1985. [3 citations pages 23, 24, et 27]

E. Hüllermeier et M. Rifqi. A fuzzy variant of the rand index for comparing clustering structures. In *13th IFSA World Congress*, 2009. [6 citations pages 17, 55, 62, 65, 67, et 97]

E. Hüllermeier, M. Rifqi, S. Henzgen, et R. Senge. Comparing Fuzzy Partitions : A Generalization of the Rand Index and Related Measures. *IEEE Transactions on Fuzzy Systems*, (20) :546–556, 2012. [9 citations pages 57, 60, 67, 101, 114, 121, 122, 124, et 126]

N. Hurley et S. Rickard. Comparing measures of sparsity. *IEEE Trans. on Information Theory*, 55(10), 2009. [2 citations pages 109 et 184]

P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37 :547–579, 1901. [3 citations pages 12, 26, et 69]

A. K. Jain. Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 30, 2009. [Cité page 11]

A. K. Jain et R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, 1988. [3 citations pages 11, 12, et 27]

A. K. Jain, M. N. Murty, et P. J. Flynn. Data clustering : A review. *ACM Comp. Surv.*, 31, 1999. [Cité page 11]

I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer, 2nd edition edition, 2002. [Cité page 83]

J. Karvanen et A. Cichocki. Measuring sparseness of noisy signals. In *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation*, 2003. [Cité page 108]

U. Kaymak et H. R. van Nauta Lemke. A sensitivity analysis approach to introducing weight factors into decision functions in fuzzy multicriteria decision making. *Fuzzy Sets and Systems*, 97(2) :169 – 182, 1998. [Cité page 69]

M. G. Kendall et A. Stuart. *The Advanced Theory of Statistics*, volume 2. Griffin, 1961. [Cité page 31]

E. P. Klement et R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005. [Cité page 44]

A. M. Krieger et P. E. Green. A generalized rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification*, 16 : 63–89, 1999. ISSN 0176-4268. [2 citations pages 1 et 12]

R. Krishnapuram et J. M. Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1(2), 1993. [5 citations pages 8, 11, 55, 147, et 148]

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 :83–97, 1955. [Cité page 36]

S. Kulczynski. Die pflanzenassociationen der pienenen. *Bulletin international de l'Académie Polonaise des Sciences et des Lettres, classe des sciences mathématiques et naturelles, Série B, Supplément II*, 2 :57–203, 1927. [3 citations pages 24, 25, et 27]

B. Larsen et C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM. [Cité page 33]

H. Le Capitaine et C. Frélicot. Classification with reject options in a logical framework : a fuzzy residual implication approach. In *13th Int. Fuzzy Systems Association World Congress*. IFSA, 2009. [Cité page 51]

H. Le Capitaine et C. Frélicot. A family of measures for best top-n class-selective decision rules. *Pattern Recognition*, 45(1) :552 – 562, 2012. [Cité page 73]

A. Lipkus. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26(1) :263–265, October 1999. [Cité page 21]

F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences (partie ii). *Etude du Centre Scientifique IBM France*, F069, 1984. [2 citations pages 31 et 61]

M. Mas, M. Monserrat, J. Torrens, et E. Trillas. A survey on fuzzy implication functions. *Trans. Fuz Sys.*, 15(6) :1107–1121, December 2007. ISSN 1063-6706. [Cité page 50]

B. H. McConnaughey. The determination and analysis of plankton communities. *Marine Research*, Special No :1–40, 1964. [2 citations pages 25 et 27]

J. McQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967. [Cité page 11]

M. Meilă. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5) :873–895, May 2007. ISSN 0047-259X. [2 citations pages 33 et 38]

M. Meilă et D. Heckerman. An experimental comparison of model-based clustering methods. *Mach. Learn.*, 42(1/2) :9–29, January 2001. ISSN 0885-6125. [3 citations pages 32, 34, et 35]

K. Menger. Statistical methods. In *Proceedings of National Academy of Science USA*, volume 28, pages 535–537, 1942. [Cité page 42]

S. Ovchinnikov. Similarity relations, fuzzy partitions, and fuzzy orderings. *Fuzzy Sets Syst.*, 40(1) :107–126, March 1991. ISSN 0165-0114. [Cité page 132]

N. R. Pal, K. Pal, J. M. Keller, et J. C. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *Trans. Fuz Sys.*, 13(4) :517–530, August 2005. ISSN 1063-6706. [2 citations pages 11 et 146]

R. Quéré et C. Frélicot. A normalized soft window-based similarity measure to extend the rand index. In *20th IEEE Int. Conf. on Fuzzy Systems*, 2011a. [9 citations pages 2, 99, 121, 122, 124, 126, 128, 129, et 131]

R. Quéré et C. Frélicot. A general framework for a class of comparison indices of soft partitions. In *14th IFSA World Congress*, 2011b. [5 citations pages 68, 121, 122, 123, et 132]

R. Quéré et C. Frélicot. A new index based on sparsity measures for comparing partitions. In *To appear in 9th Int. Workshop in Statistical Techniques in Pattern Recognition*, 2012. [6 citations pages 2, 113, 126, 128, 129, et 132]

R. Quéré, H. Le Capitaine, N. Fraisseix, et C. Frélicot. On normalizing fuzzy coincidence matrices to compare fuzzy and/or possibilistic partitions with the rand index. In *10th. IEEE Int. Conf. on Data Mining*, 2010. [9 citations pages 2, 86, 121, 122, 124, 126, 128, 129, et 131]

W. M. Rand. Objective criteria for the evaluation of clustering methods. *J. of the American Statistical Association*, 66, 1971. [6 citations pages 12, 13, 16, 21, 26, et 38]

S. Régnier. Quelques aspects mathématiques des problèmes de classification automatique. *I.C.C. Bulletin*, (4), 1965. [Réédité dans *Mathématiques et Sciences humaines* 82 :13-29, 1983]. [Cité page 34]

M. Rifqi, V. Berger, et B. Bouchon-Meunier. Discrimination power of measures of comparison. *Fuzzy Sets and Systems*, 110(2) :189 – 196, 2000. ISSN 0165-0114. [Cité page 52]

T. A. Runkler. Comparing partitions by subset similarities. In *Proceedings of the Computational intelligence for knowledge-based systems design, and 13th international conference on Information processing and management of uncertainty*, IPMU'10, pages 29–38, Berlin, Heidelberg, 2010. Springer-Verlag. [Cité page 70]

P. F. Russel et T. R. Rao. On habitat and association of species of anopheline larvae in south-eastern madras. *J. Malaria Institute India*, 3 :153–178, 1940. [Cité page 26]

B. Schweizer et A. Sklar. *Probabilistic Metric Spaces*. North-Holland, Amsterdam, 1983. [Cité page 42]

R. R. Sokal et C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38 :1409–1438, 1958. [Cité page 12]

R. R. Sokal et P. H. A. Sneath. *Principles of Numerical Taxonomy*. Freeman, San Francisco, USA, 1963. [2 citations pages 26 et 27]

F. Tavares Pereira, J. Rui Figueira, V. Mousseau, et B. Roy. Comparing two territory partitions in districting problems : Indices and practical issues. *Socio-Economic Planning Sciences*, 43(1) :72 – 88, 2009. ISSN 0038-0121. [Cité page 12]

A. Tversky. Features of Similarity. In *Psychological Review*, volume 84, pages 327–352, 1977. [2 citations pages 32 et 52]

R. Unnikrishnan, C. Pantofaru, et M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6) :929–944, June 2007. ISSN 0162-8828. [Cité page 1]

D. L. Wallace. Comment. *Journal of the American Statistical Association*, 78 :569–579, 1983. [6 citations pages 12, 17, 22, 23, 25, et 27]

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8 :338–353, 1965. [Cité page 9]

L. A. Zadeh. Similarity relations and fuzzy orderings. *Information Sciences*, 3(2) :177 – 200, 1971. ISSN 0020-0255. [2 citations pages 16 et 125]

# ANNEXE A :

# Algorithmes de clustering

## A.1 C-Moyennes Floues

L'algorithme des C-Moyennes Floues, ou Fuzzy C-Means (FCM) [Bezdek, 1981], est un algorithme de classification non-supervisée orienté prototypes. Il repose sur la résolution du problème d'optimisation suivant :

$$\min_{(U,V)} \{ \; J_m(U,V;X) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \; || \; \mathbf{x}_k - \mathbf{v}_i \; ||^2 \; \}, \tag{A.1}$$

minimisant la distance intra-clusters pour produire une $c$-partition floue $U^\star$ de $X$ et un ensemble $V^\star = \{v_1^\star, \; v_2^\star, ..., v_c^\star\}$ de $c$ prototypes centroïdes associés à chaque cluster. Le nombre $c$ de clusters est fixé par l'utilisateur, de même que le paramètre $m > 1$, appelé *exposant de flou*, qui détermine le degré de flou de la partition $U^\star$. Plus $m$ est petit et plus la partition sera stricte ; plus $m$ est grand et plus les frontières de ses clusters seront douces.

La minimisation de la fonction objectif $J_m$ s'effectue sous les contraintes :

$$0 < \sum_{k=1}^{n} u_{ik} < n, \tag{A.2}$$

$$\sum_{i=1}^{c} u_{ik} = 1, \tag{A.3}$$

cette dernière découlant de la normalisation imposée aux vecteurs d'appartenance qui composent les partitions de $\mathbb{M}_{fcn}$. Le couple $(U^\star, V^\star)$ est alors obtenu par itérations successives des formules de mise à jour suivantes :

$$u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{|| \; \mathbf{x}_k - \mathbf{v}_i \; ||}{|| \; \mathbf{x}_k - \mathbf{v}_j \; ||} \right)^{2/(m-1)} \right)^{-1} \tag{A.4}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{A.5}$$

jusqu'à convergence de $J_m(U, V; X)$, tel que le décrit l'algorithme 1.

**Algorithme 1:** C-Moyennes Floues

**Entrées** :

- $X$ : jeu de données
- $c$ : nombre de clusters
- $\epsilon$ : critère de terminaison

**Sorties** :

- $U^\star$ : $c$-partition floue de $X$
- $V^\star$ : $c$ prototypes

```
1 début
2     Initialisation aléatoire de la partition floue U_t
3     répéter
4         Mise à jour de V_t selon (A.5)
5         Mise à jour de U_t selon (A.4)
6     jusqu'à || U_t − U_{t−1} ||< ε
7     U* ⟵ U_t
8     V* ⟵ U_t
9 fin
```

Il est important de noter que le couple $(U^\star, V^\star)$ produit par FCM est un optimum local de $J_m$ et dépend ainsi grandement de l'initialisation de la partition $U_t$ (ligne 2). Par ailleurs, on trouve dans la littérature [Bezdek, 1981] une formulation équivalente de cet algorithme dans laquelle ce sont prototypes $V_t$ qui sont initialisés (il faut alors échanger les lignes 4 et 5). Enfin, et bien que cela survient très rarement en pratique, l'utilisateur devra prendre garde à ce que jamais les prototypes $\mathbf{v}_i$ ne se confondent avec l'un des individus de $X$, sous peine de ne pouvoir mettre à jour les degrés d'appartenance selon (A.4). Pour plus de détails, le lecteur pourra se référer aux travaux de Pal et al. [2005].

## A.2 C-Moyennes Possibilistes

L'algorithme des C-Moyennes Possibilistes, ou Possibilistic C-Means (PCM) [Krishnapuram et Keller, 1993], est un algorithme de classification non-supervisée orienté prototypes très analogue à l'algorithme des C-Moyennes Floues, et il ne diffère de ce dernier dans sa construction que par sa fonction objectif. Il repose ainsi sur le problème d'optimisation suivant :

$$\min_{(T,V)} \Big\{ P_m(T,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(t_{ik})^m \,||\, \mathbf{x}_k - \mathbf{v}_i \,||^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n}(1-t_{ik})^m \Big\}, \tag{A.6}$$

où $T$ est une $c$-partition possibiliste, $V$ est un ensemble de $c$ prototypes centroïdes , $m$ est l'exposant de flou qui détermine le degré de flou de la partition $T^\star$ optimale résultant de la minimisation de $P_m$ et où les $\{\gamma_i\}$ sont des termes de pénalisation défini par l'utilisateur, pour $1 \leq i \leq c$. Contrairement à FCM, l'algorithme PCM produit en sortie une partition de $\mathbb{M}_{pcn}$ dont les degrés d'appartenance ne sont plus relatifs mais absolus et pallie ainsi certains problèmes liés à la gestion des points aberrants inhérents à FCM, comme discuté plus tôt dans ce mémoire, page 8.

Tout comme pour l'algorithme FCM, la minimisation de la fonction objectif $P_m$ s'effectue par itérations successives des formules de mise à jour suivantes :

$$t_{ik} = \frac{1}{1 + \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||^2}{\gamma_i}\right)^{1/(m-1)}}, \tag{A.7}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} t_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{n} t_{ik}^m}, \tag{A.8}$$

jusqu'à convergence de $P_m$. Le pseudocode de PCM est donné par l'algorithme 2.

**Algorithme 2:** C-Moyennes Possibilistes

**Entrées :**

- $X$ : jeu de données
- $c$ : nombre de clusters
- $\epsilon$ : critère de terminaison

**Sorties :**

- $T^\star$ : $c$-partition possibiliste de $X$
- $V^\star$ : $c$ prototypes

```
début
    Initialisation de la partition possibiliste T_t
    répéter
        Mise à jour de V_t selon (A.8)
        Mise à jour de T_t selon (A.7)
    jusqu'à || U_t − U_{t−1} ||< ϵ
    T* ⟵ T_t
    V* ⟵ U_t
fin
```

Bien que le choix des coefficients $\{\gamma_i\}$ soit laissé à l'utilisateur, Krishnapuram et Keller [1993] préconisent de les calculer selon :

$$\gamma_i = K \frac{\sum_{k=1}^{n} u_{ik}^m \,||\, \mathbf{x}_k - \mathbf{v}_i \,||^2}{\sum_{k=1}^{n} u_{ik}^m}, \; K > 0 \tag{A.9}$$

où les termes $\{u_{ik}\}$ sont les degrés d'appartenance de la partition $U^\star$ de $X$ terminale produite par FCM et $K$ est fixé par l'utilisateur. De la même manière, les auteurs suggèrent d'initialiser la partition $T_t$ selon cette partition $U^\star$.

Pour générer des partitions possibilistes de $\mathbb{M}_{pcn}^{\leq}$, nous proposons de minimiser la fonction objectif de PCM sous la contrainte supplémentaire suivante :

$$\sum_{i=1}^{c} t_{ik} = b_k \tag{A.10}$$

où les coefficients $\{b_k\}$ sont choisis inférieurs ou égaux à 1. Nous suggérons de les définir tels que $b_k = u_{(1)k}$ pour $1 \leq k \leq n$, avec $u_{(1)k}$ le plus grand degré d'appartenance de $\mathbf{x}_k$ à l'un des clusters de la partition $U^\star$ issue de FCM. Pour $m = 2$, la formule de mise à jour de la partition $T$ devient alors :

$$t_{ik} = \frac{b_k + \sum_{j=1}^{c} \frac{\gamma_i - \gamma_j}{||\mathbf{x}_k - \mathbf{v}_j||^2 + \gamma_j}}{\frac{||\mathbf{x}_k - \mathbf{v}_i||^2 + \gamma_i}{\sum_{j=1}^{c} ||\mathbf{x}_k - \mathbf{v}_j||^2 + \gamma_j}} \tag{A.11}$$

que l'on pourra substituer à (A.7) dans l'algorithme 2.

# ANNEXE B :

# Résultats complémentaires

Nous donnons dans cette annexe quelques résultats complémentaires obtenus sur les expérimentations décrites au Chapitre 4, page 79. Certains résultats obtenus avec les indices auxquels nous confrontons nos propositions, donnés au Chapitre 4, sont repris pour faciliter la comparaison.

## B.1 Une extension de mesures strictes

### B.1.1 Partitions non-strictes synthétiques (E1)

FIGURE B.1 – Mosaïques des valeurs de comparaison obtenues avec les extensions de $RI$ et de $JI$ de Anderson et al. pour les partitions floues de $S$ et possibilistes de $S'$

FIGURE B.2 – Mosaïques des valeurs de comparaison obtenues avec les extensions de $RI$ et $JI$ proposées par Campello, Borgelt et Quéré et al., pour $\top_P$ et $\top_S$. Première ligne : partitions floues de $S$. Deuxième ligne : partitions possibilistes de $S'$.

FIGURE B.3 – Mosaïques des valeurs de comparaison obtenues avec les extensions de $RI$ et $JI$ proposées par Campello, Borgelt et Quéré et al., pour $\top_{H_0}$ et $\top_{H_{100}}$. Première et deuxième ligne : partitions floues de $\mathcal{S}$. Troisième et quatrième ligne : partitions possibilistes de $\mathcal{S}'$.

FIGURE B.4 – Mosaïques des valeurs de comparaison obtenues avec les extensions de $RI$ et $JI$ proposées par Campello, Borgelt et Quéré et al., pour $\top_{F_{0.1}}$ et $\top_{H_{100}}$. Première et deuxième ligne : partitions floues de $\mathcal{S}$. Troisième et quatrième ligne : partitions possibilistes de $\mathcal{S}'$.

### B.1.2 Données synthétiques - Partitions non strictes (E3)

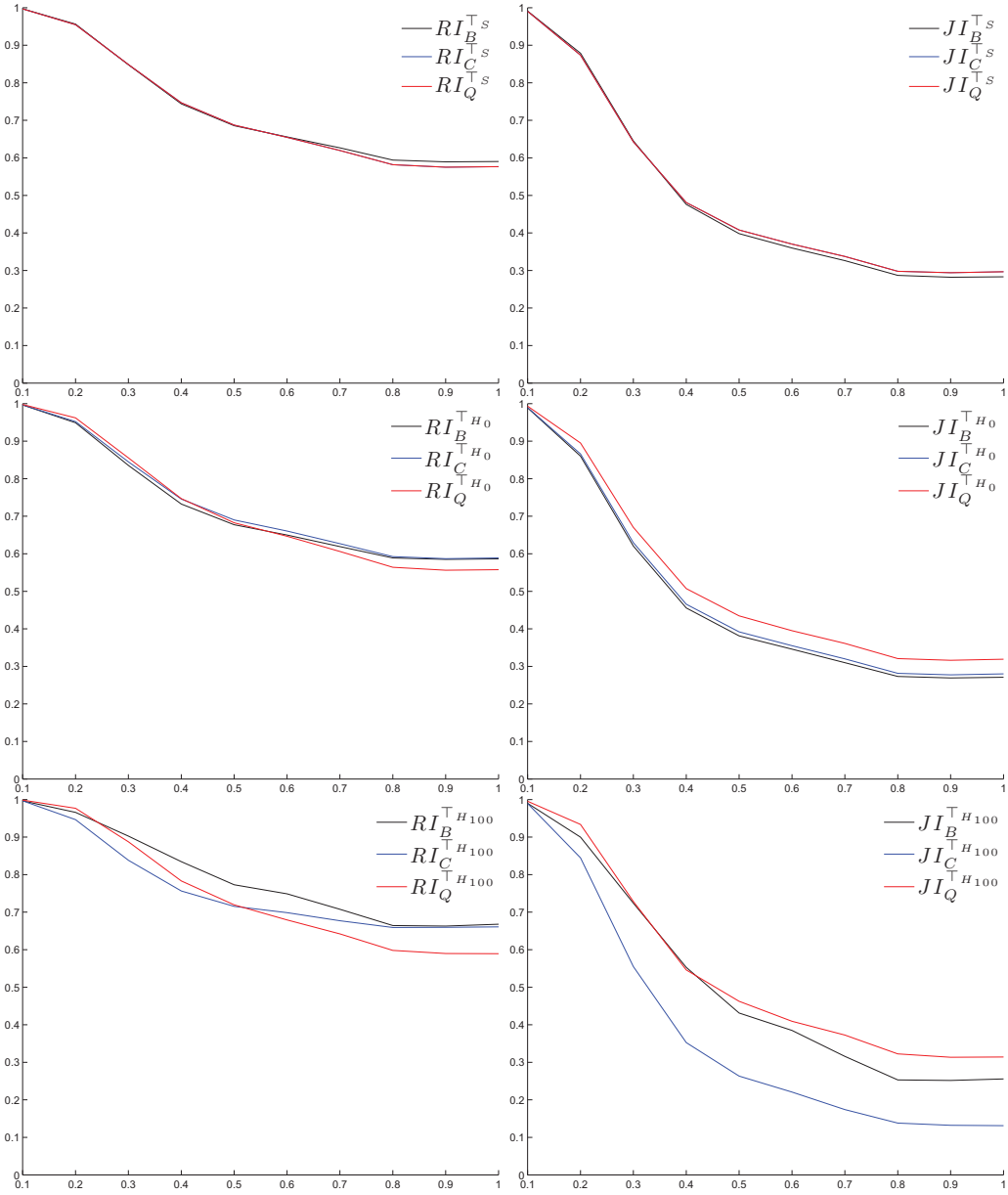



FIGURE B.5 – Comparaison d'une collection de partitions floues $V_\sigma$ à une partition stricte de référence $U_{k^\star}$ avec les indices de Rand et de Jaccard dérivés selon Anderson et al., Campello, Borgelt et notre proposition avec $\top_S$, $\top_{H_0}$ et $\top_{H_{100}}$.

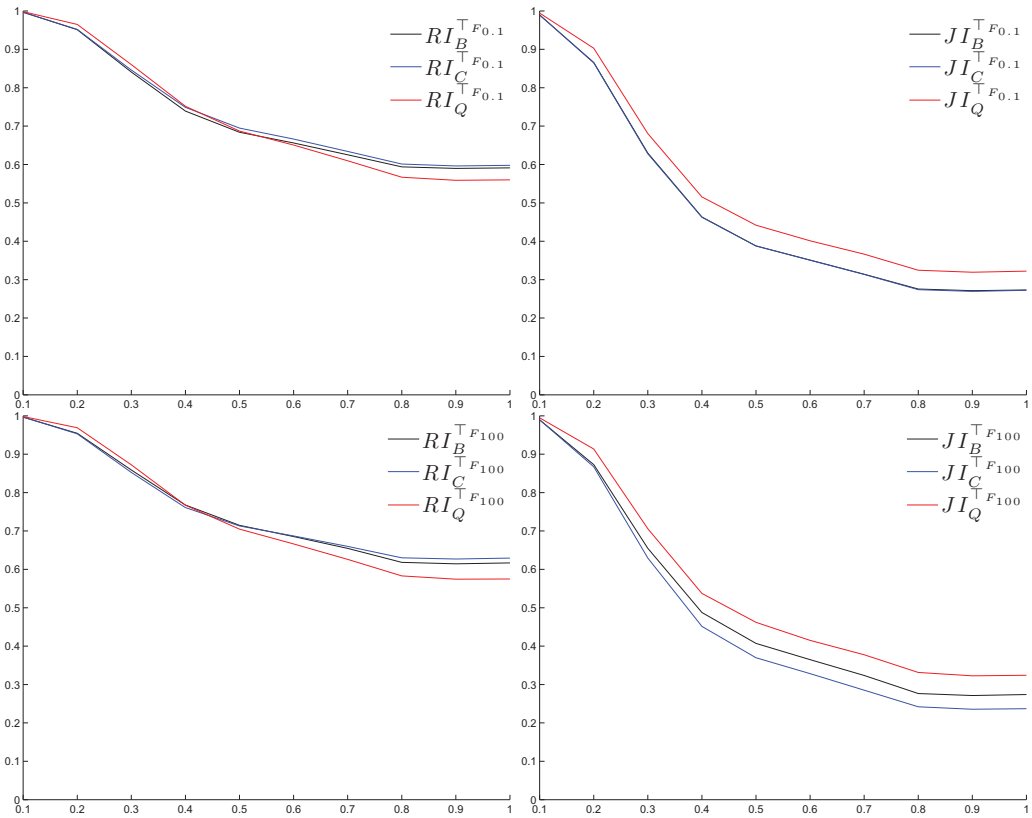



FIGURE B.6 – Comparaison d'une collection de partitions floues $V_\sigma$ à une partition stricte de référence $U_{k^\star}$ avec les indices de Rand et de Jaccard dérivés selon Campello, Borgelt et notre proposition avec $\top_{F_{0.1}}$ et $\top_{F_{100}}$.

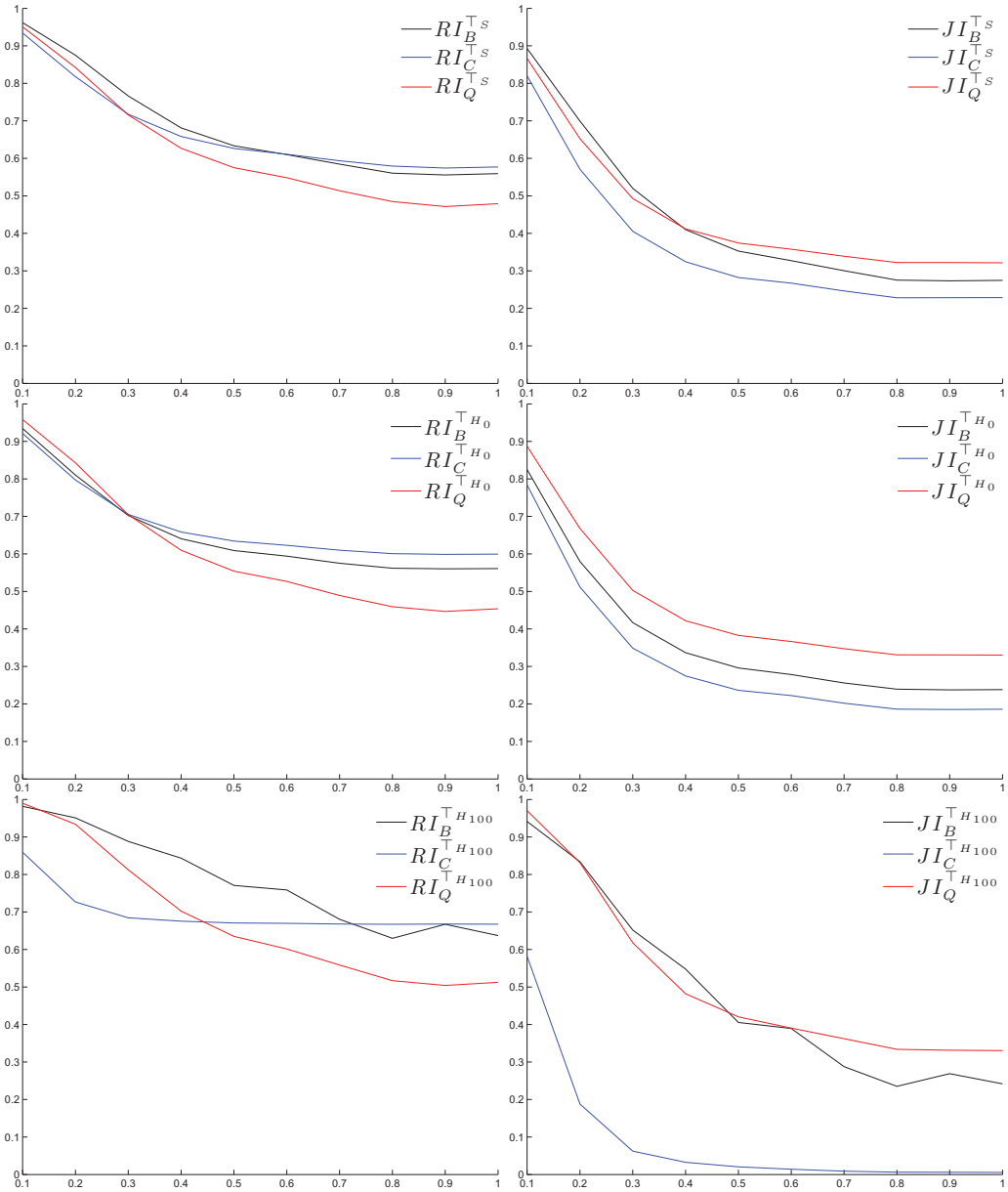



FIGURE B.7 – Comparaison d'une collection de partitions possibilistes $V_\sigma$ à une partition stricte de référence $U_{k^\star}$ avec les indices de Rand et de Jaccard dérivés selon Anderson et al., Campello, Borgelt et notre proposition avec $\top_S$, $\top_{H_0}$ et $\top_{H_{100}}$.

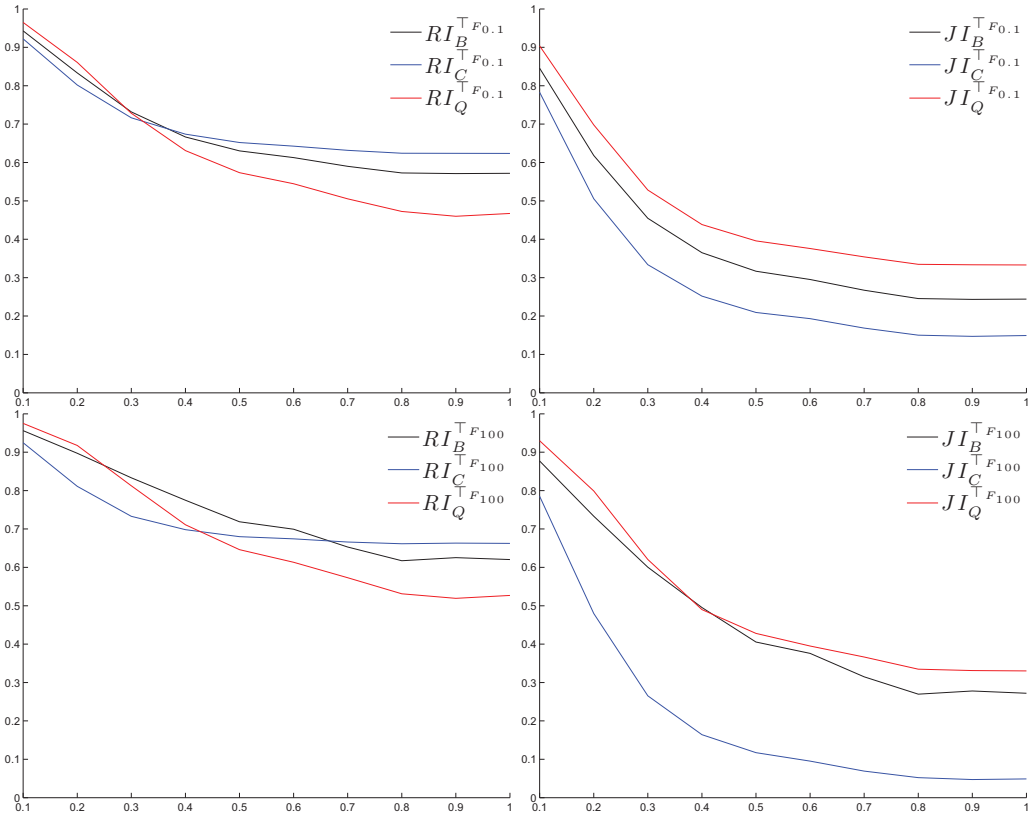



FIGURE B.8 – Comparaison d'une collection de partitions possibilistes $V_\sigma$ à une partition stricte de référence $U_{k^\star}$ avec les indices de Rand et de Jaccard dérivés selon Campello, Borgelt et notre proposition avec $\top_{F_{0.1}}$ et $\top_{F_{100}}$.

#### B.1.2.0.a Partitions non-strictes de données réelles (E4)

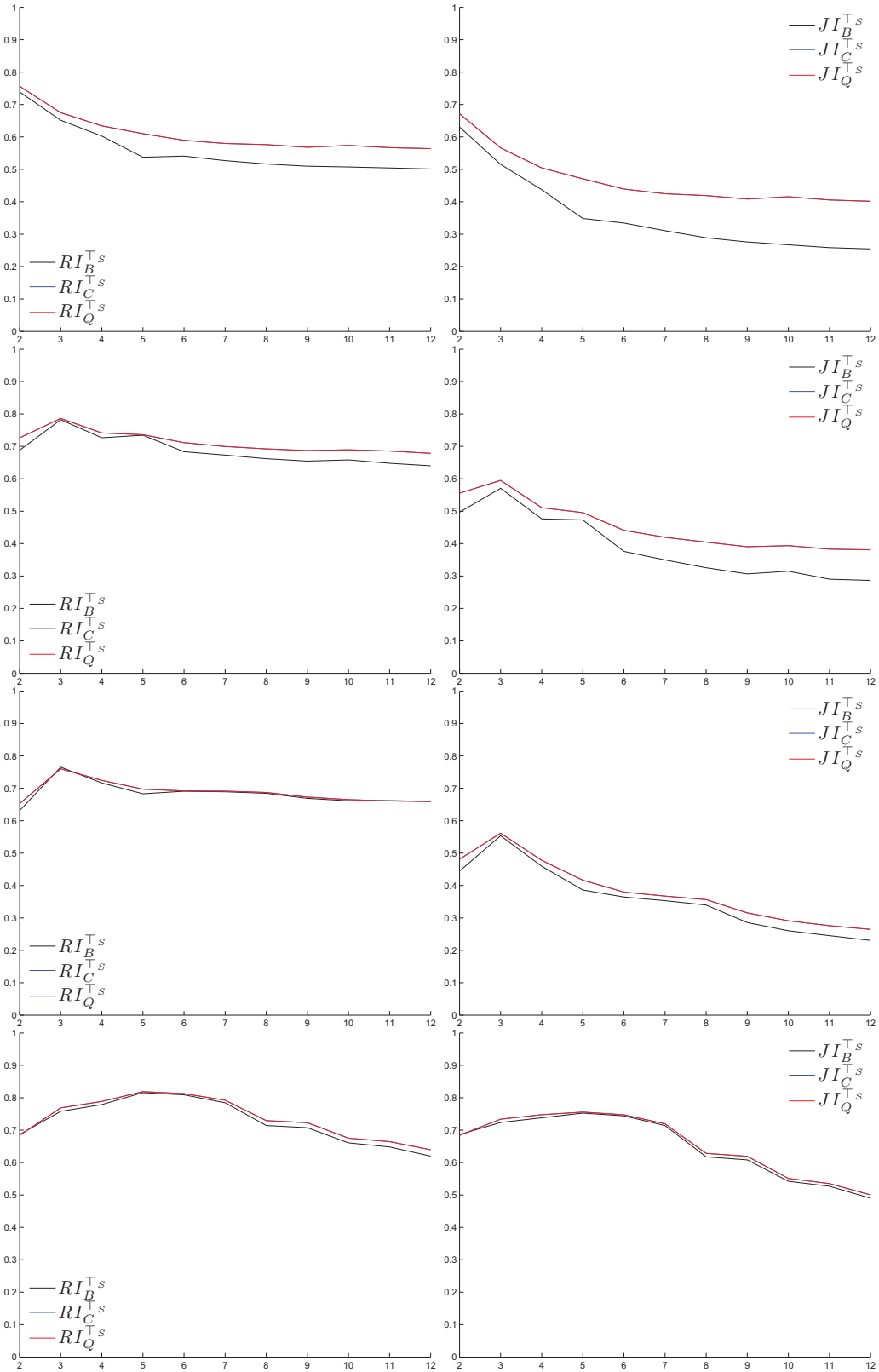



FIGURE B.9 – Valeurs des indices pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

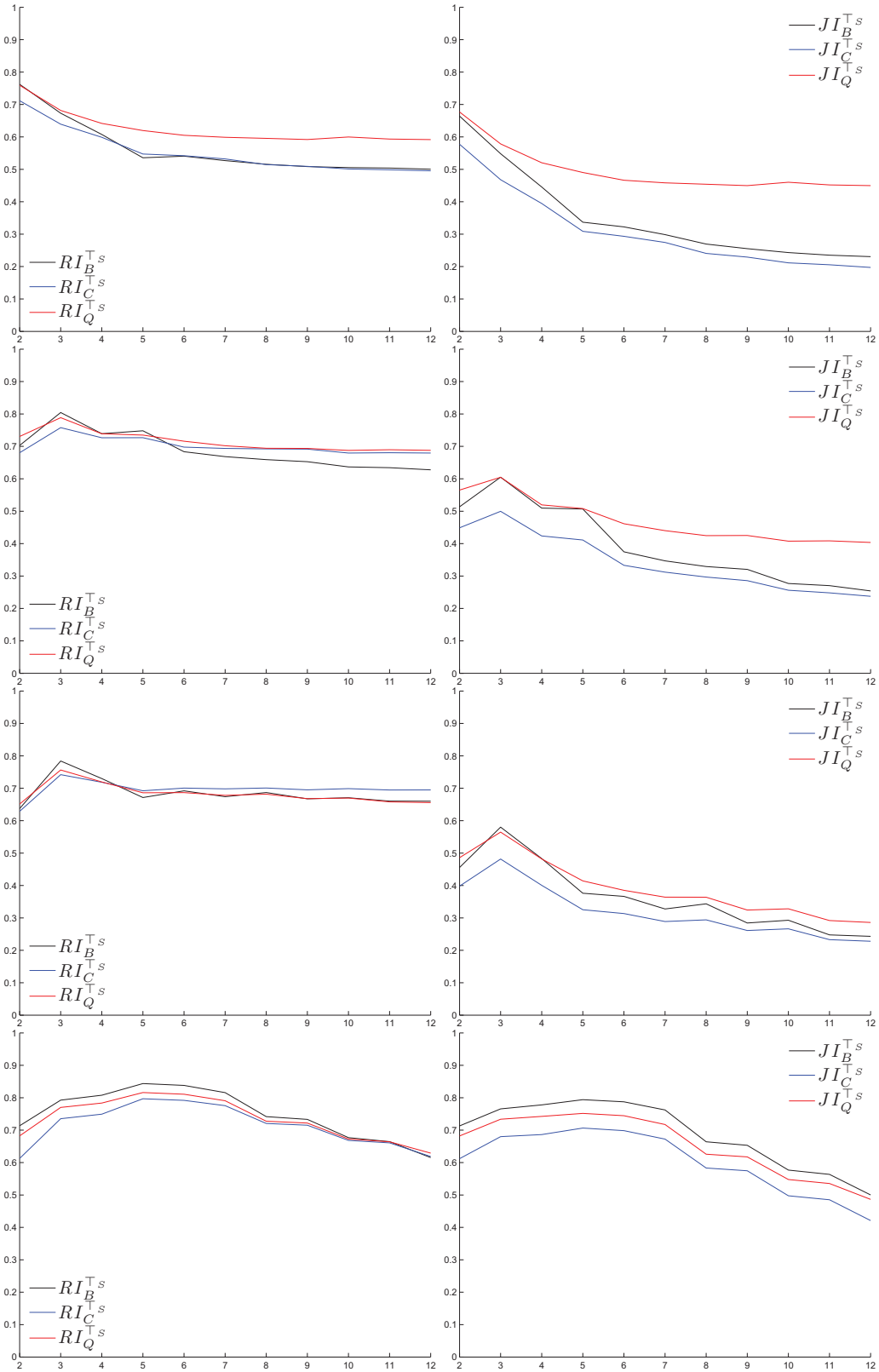



FIGURE B.10 – Valeurs des indices pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

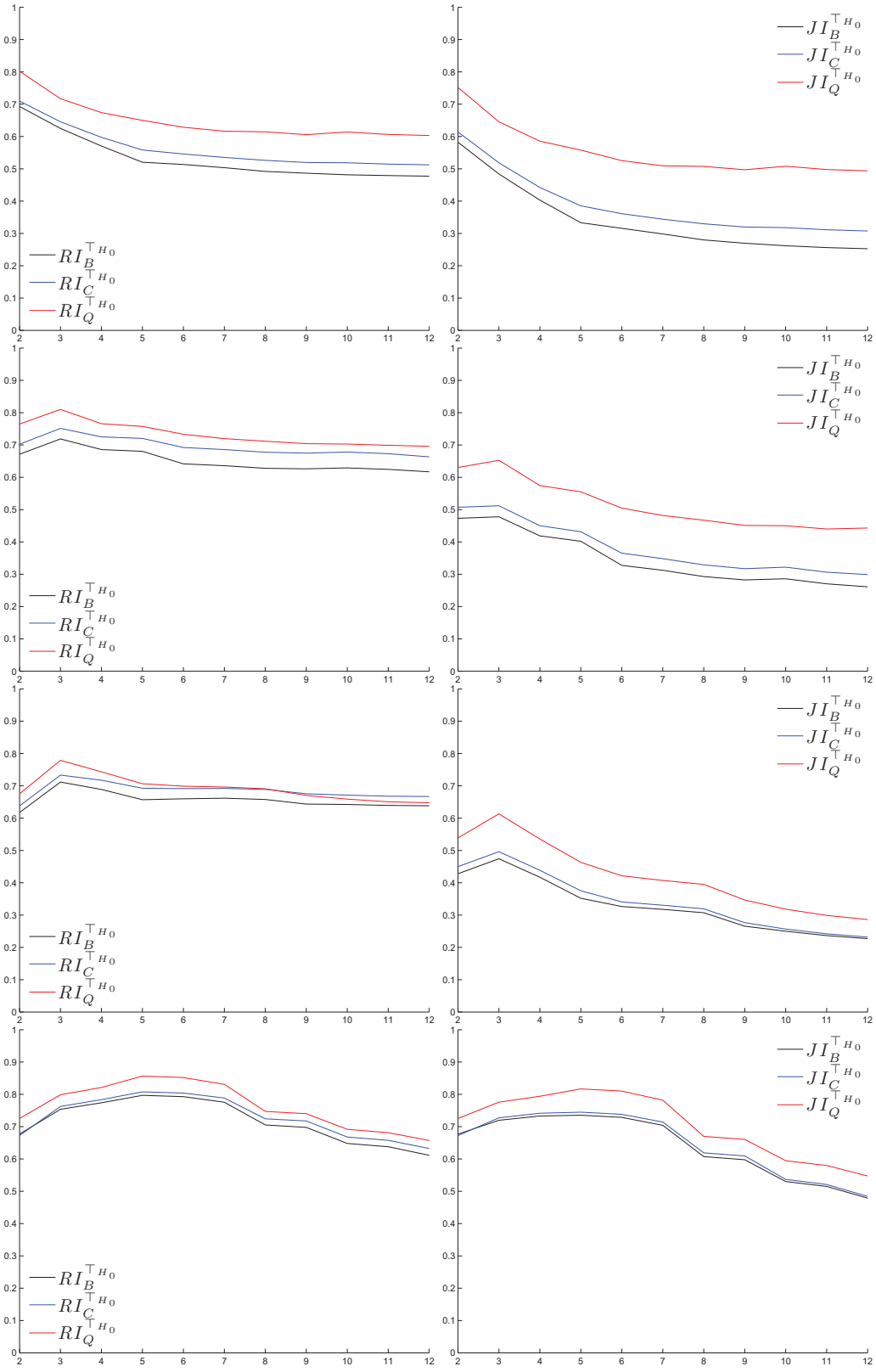



FIGURE B.11 – Valeurs des indices pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

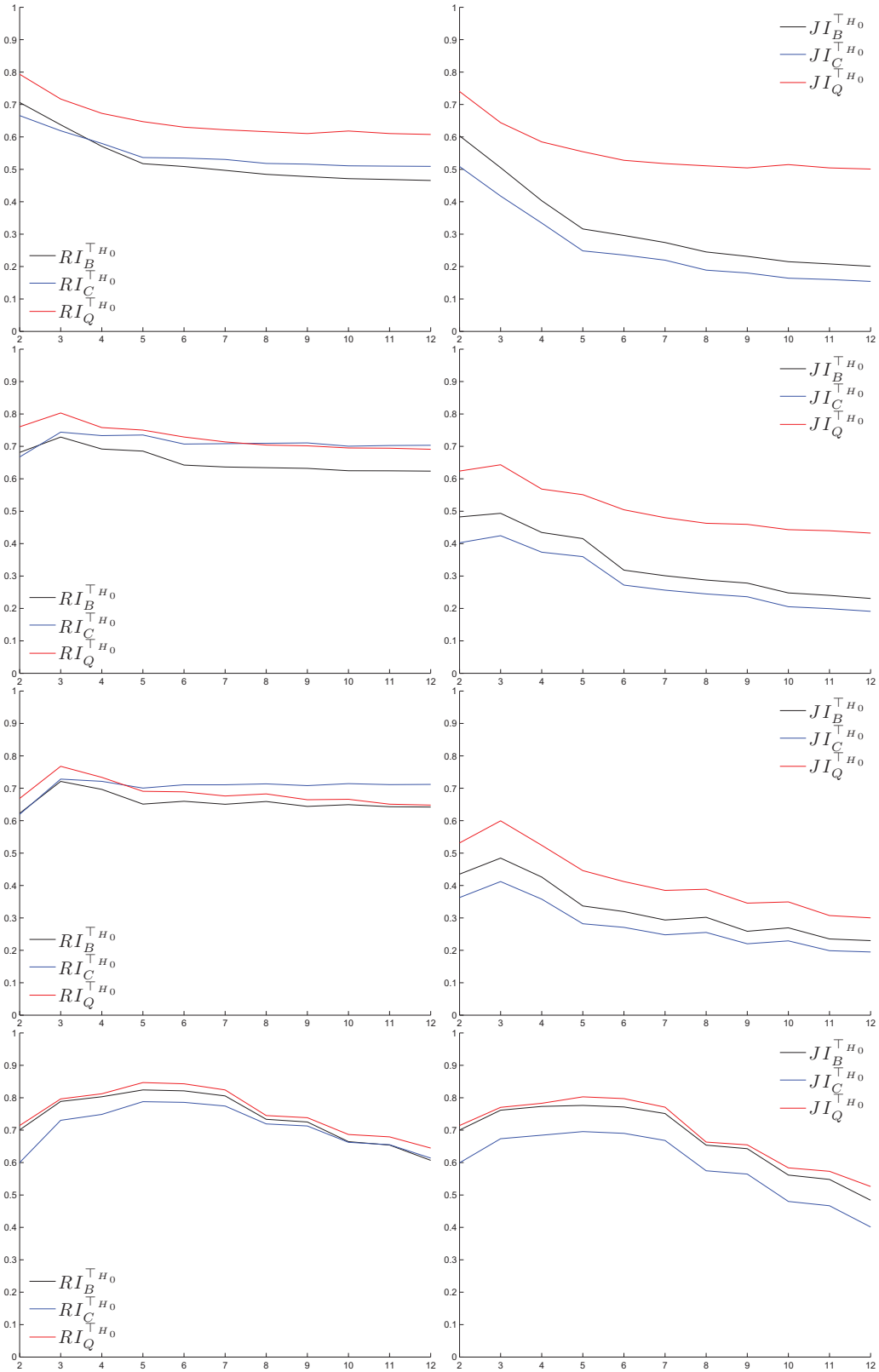



FIGURE B.12 – Valeurs des indices pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

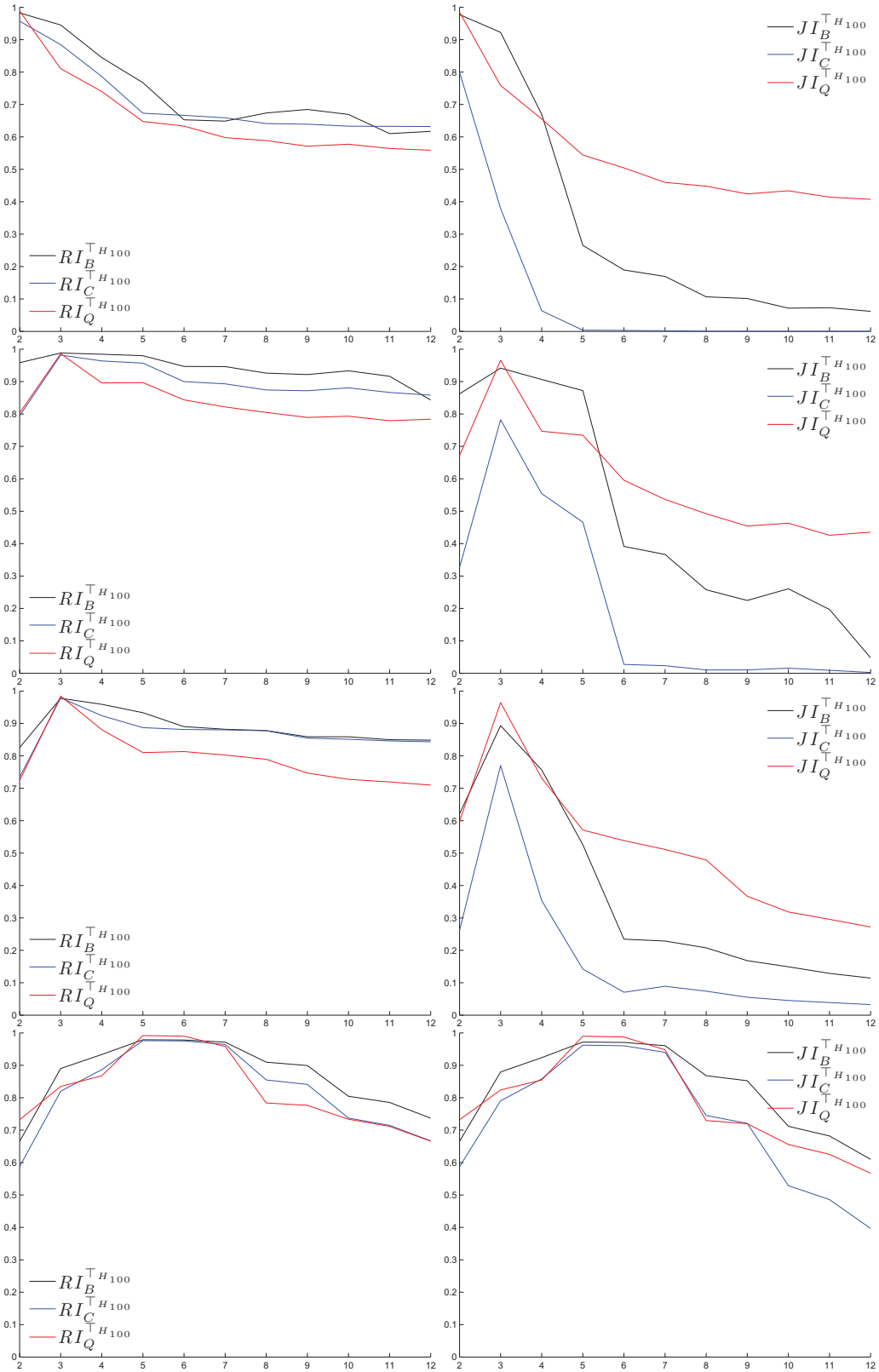



FIGURE B.13 – Valeurs des indices pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

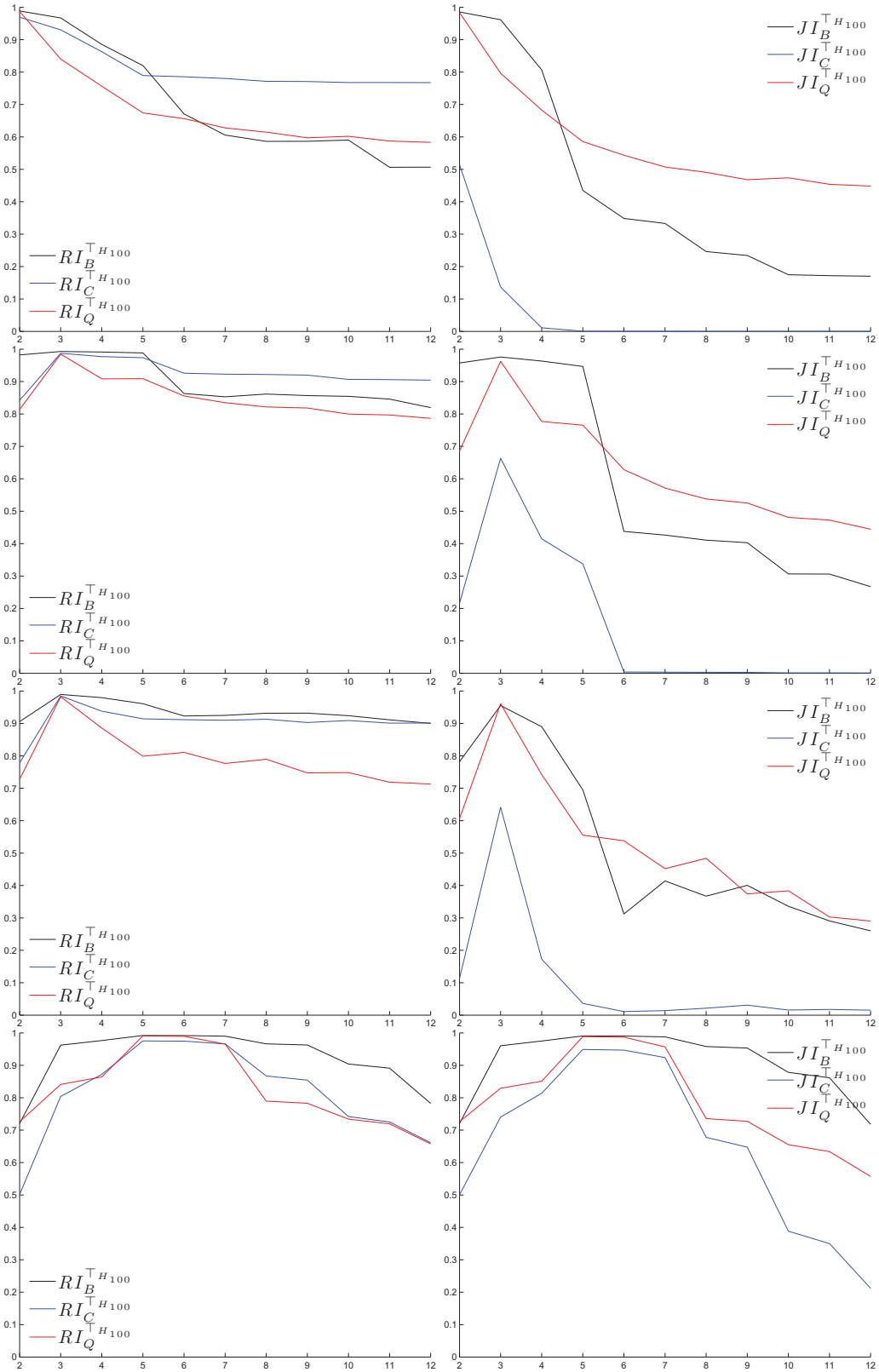



FIGURE B.14 – Valeurs des indices pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

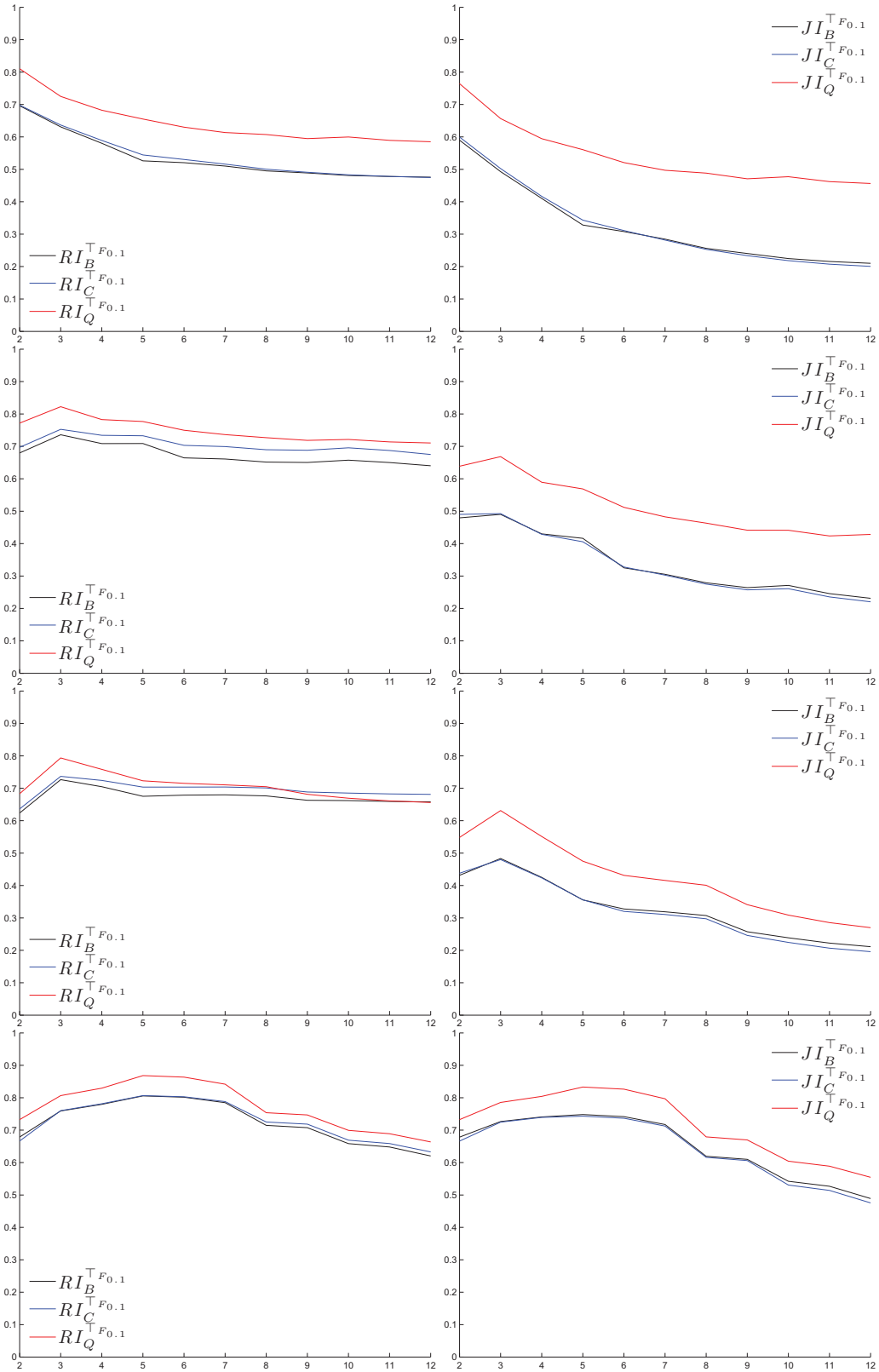



FIGURE B.15 – Valeurs des indices pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

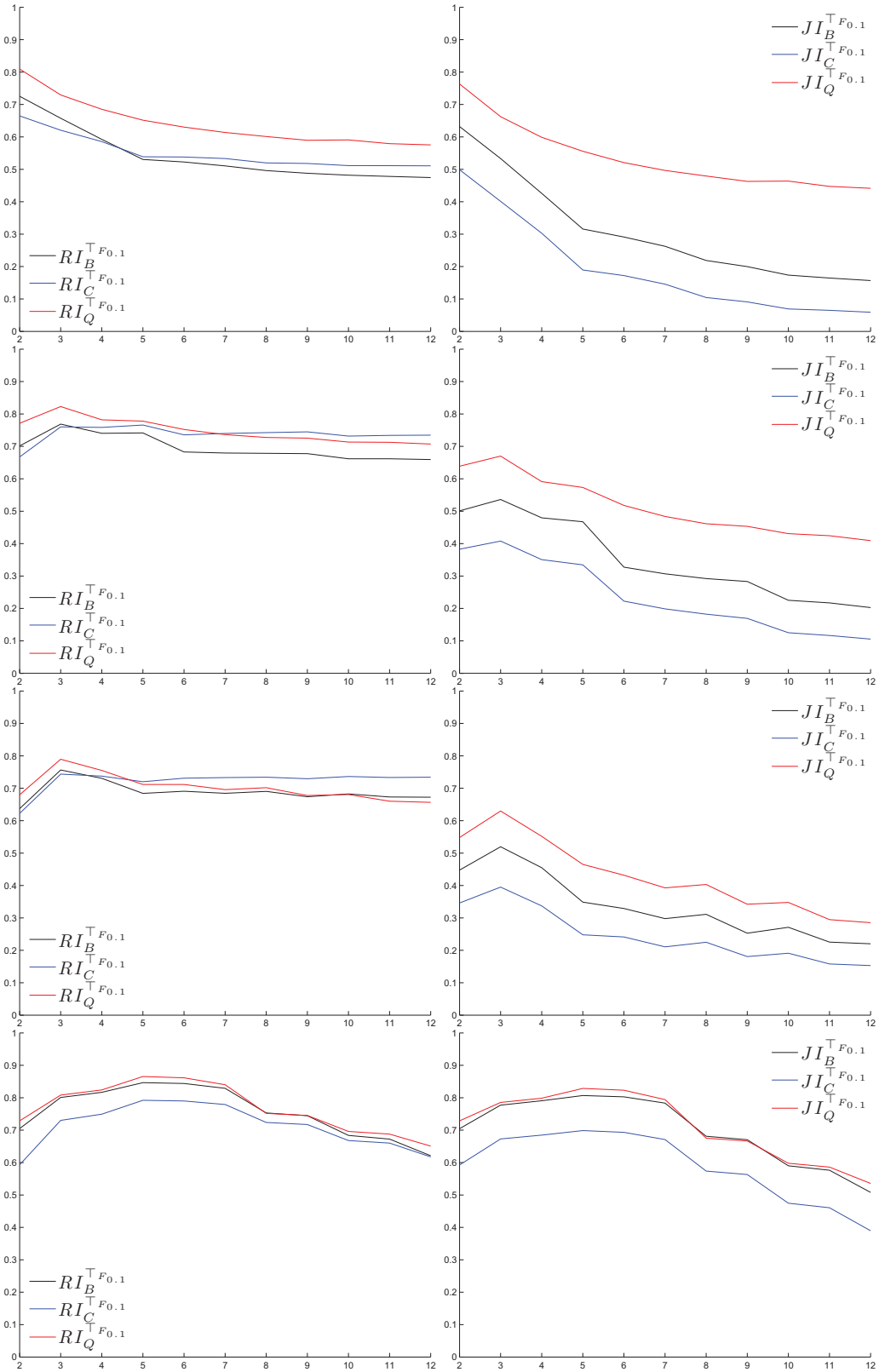



FIGURE B.16 – Valeurs des indices pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

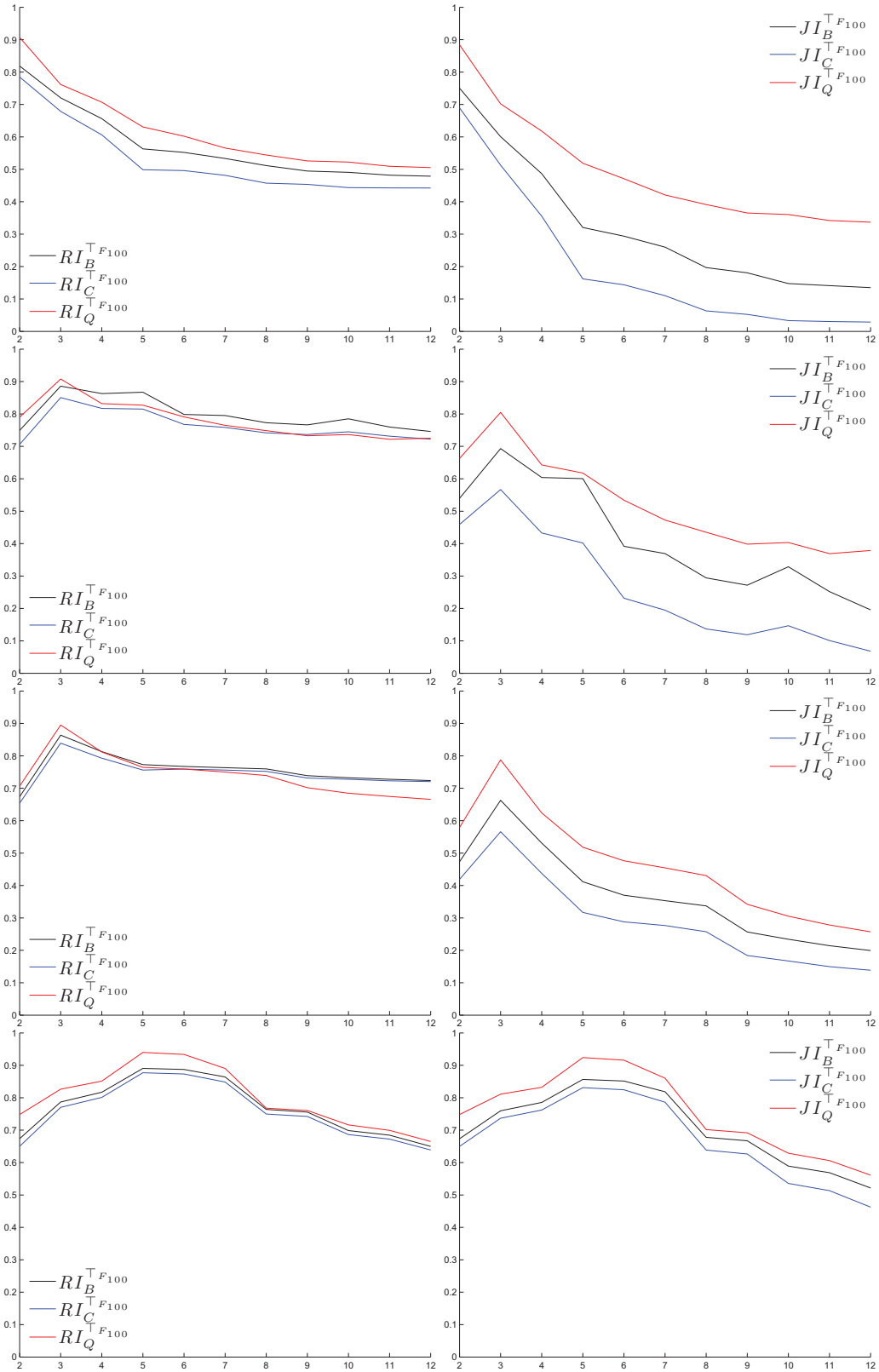



FIGURE B.17 – Valeurs des indices pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

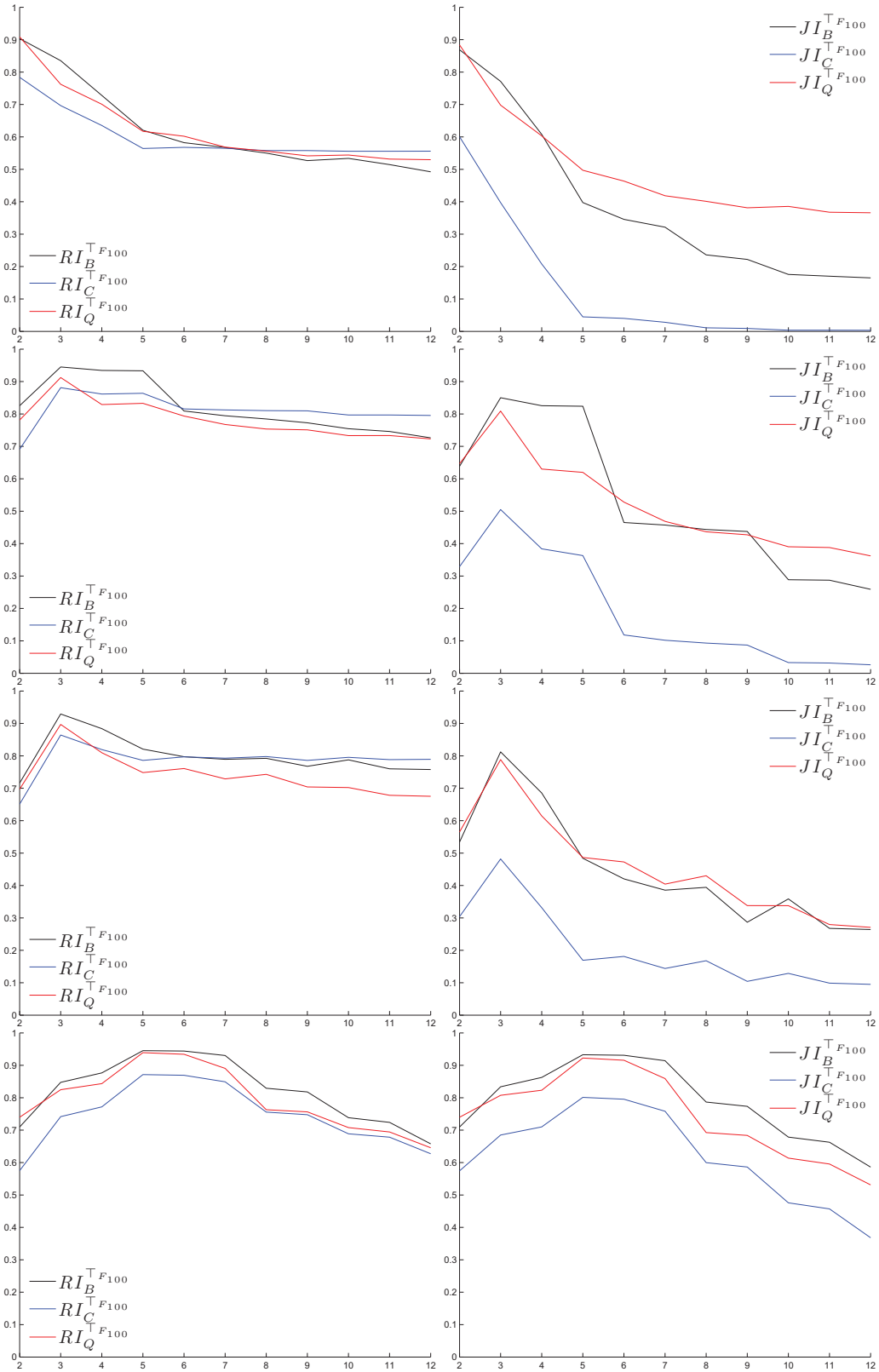



FIGURE B.18 – Valeurs des indices pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

## B.2 Une mesure native orientée individus

Sont donnés ici des résultats complémentaires illustrant l'influence de la norme triangulaire sur notre indice $RI^{\top}_{NSW}$, présenté page 97. Par ailleurs, ces résultats illustrent aussi l'utilisation de la fenêtre d'Epanechnikov $w_E$. La Table B.1 donne les paramétrages utilisés.

TABLE B.1 – Paramétrages utilisés pour illustrer l'influence de $\top$ sur $RI^{\top}_{NSW}$.

| Paramétrage | $\top$ | $w$ | $t$ | $r$ | $o$ |
|---|---|---|---|---|---|
| (g) | $\top_S$ | $w_E$ | $\frac{\sqrt{2}}{2}$ | 0 | × |
| (h) | $\top_{H_0}$ | $w_E$ | $\frac{\sqrt{2}}{2}$ | 0 | × |
| (i) | $\top_{H_{100}}$ | $w_E$ | $\frac{\sqrt{2}}{2}$ | 0 | × |
| (j) | $\top_{F_{0.1}}$ | $w_E$ | $\frac{\sqrt{2}}{2}$ | 0 | × |
| (k) | $\top_{F_{100}}$ | $w_E$ | $\frac{\sqrt{2}}{2}$ | 0 | × |

### B.2.1 Partitions non-strictes synthétiques (E1)

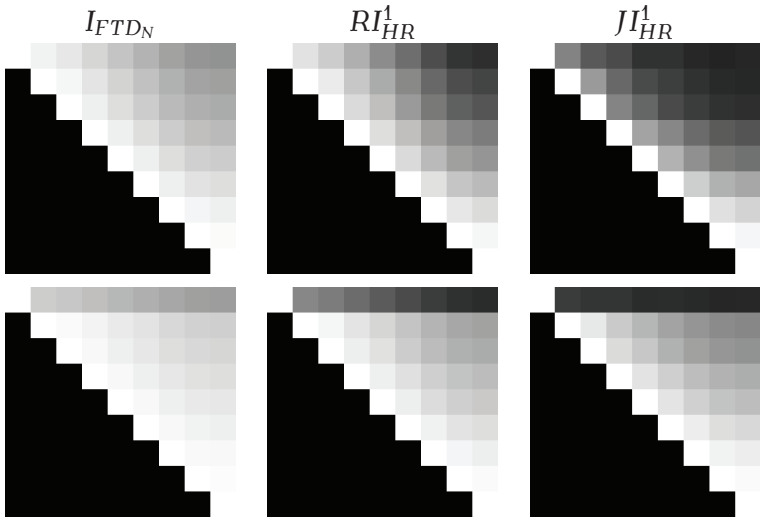



FIGURE B.19 – Mosaïques des valeurs de comparaison obtenues avec $I_{FTD_N}$, $RI^1_{HR}$ et $JI^1_{HR}$. Première ligne : partitions floues de $\mathcal{S}$. Deuxième ligne : partitions possibilistes de $\mathcal{S}'$.

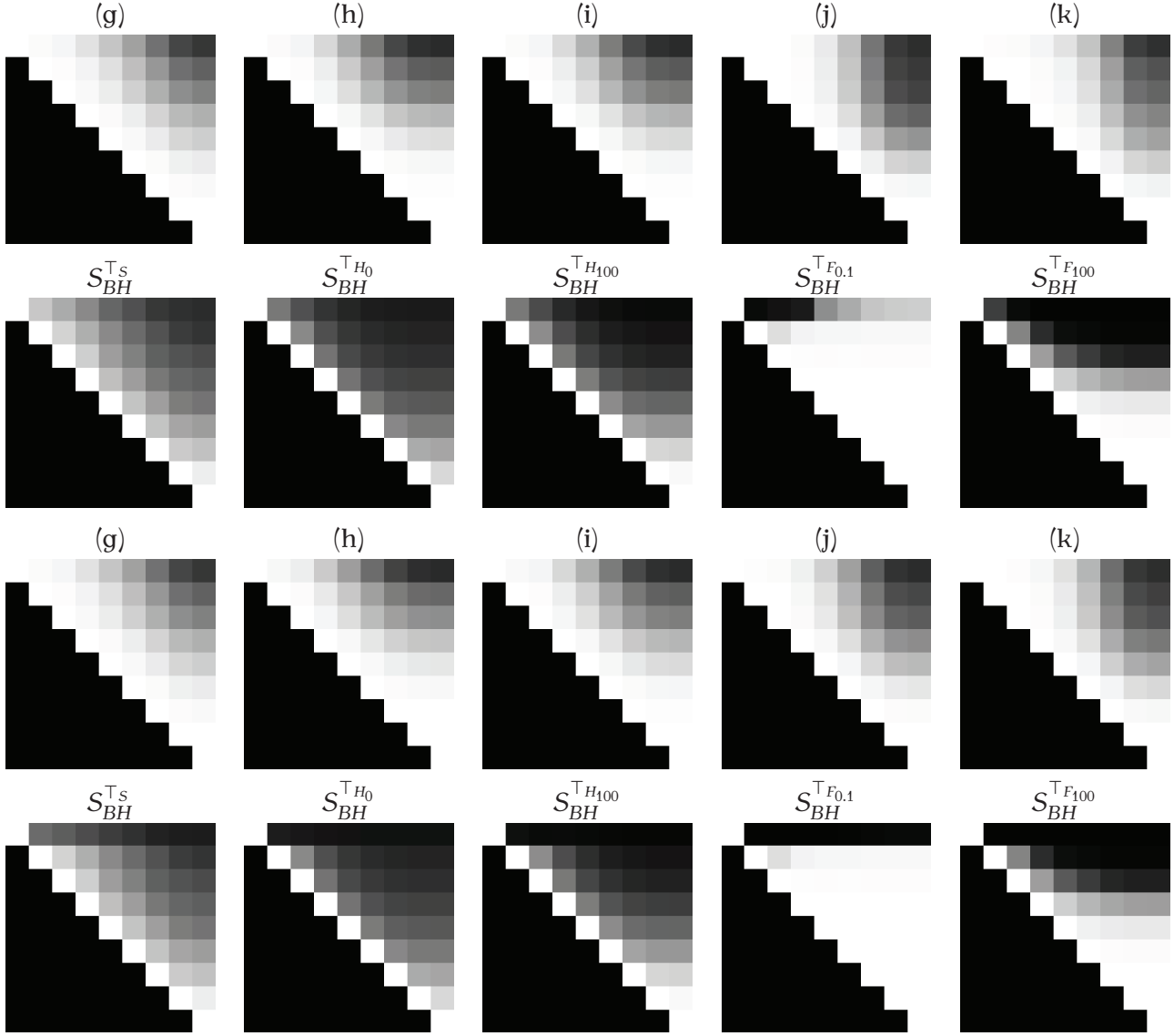



FIGURE B.20 – Mosaïques des valeurs de comparaison obtenues avec $RI_{NSW}^{\top}$ et $S_{BH}^{\top}$ selon les six paramétrages donnés par la Table B.1 pour les partitions floues de $\mathcal{S}$. Première et deuxième lignes : partitions floues de $\mathcal{S}$. Troisième et quatrième ligne : partitions possibilistes de $\mathcal{S}'$.

### B.2.2 Données synthétiques - Partitions non strictes (E3)

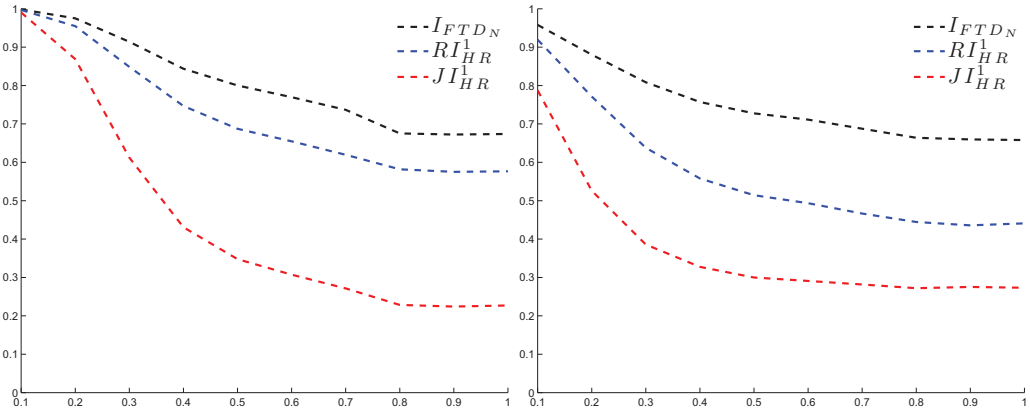



FIGURE B.21 – Comparaison de deux collections de partitions non strictes à une partition stricte de référence $U_{k^\star}$ avec $RI^1_H$, $JI^1_H$ et $I_{FTD_N}$. À gauche : partitions floues $V_\sigma$. À droite : partitions possibilistes $V'_\sigma$.

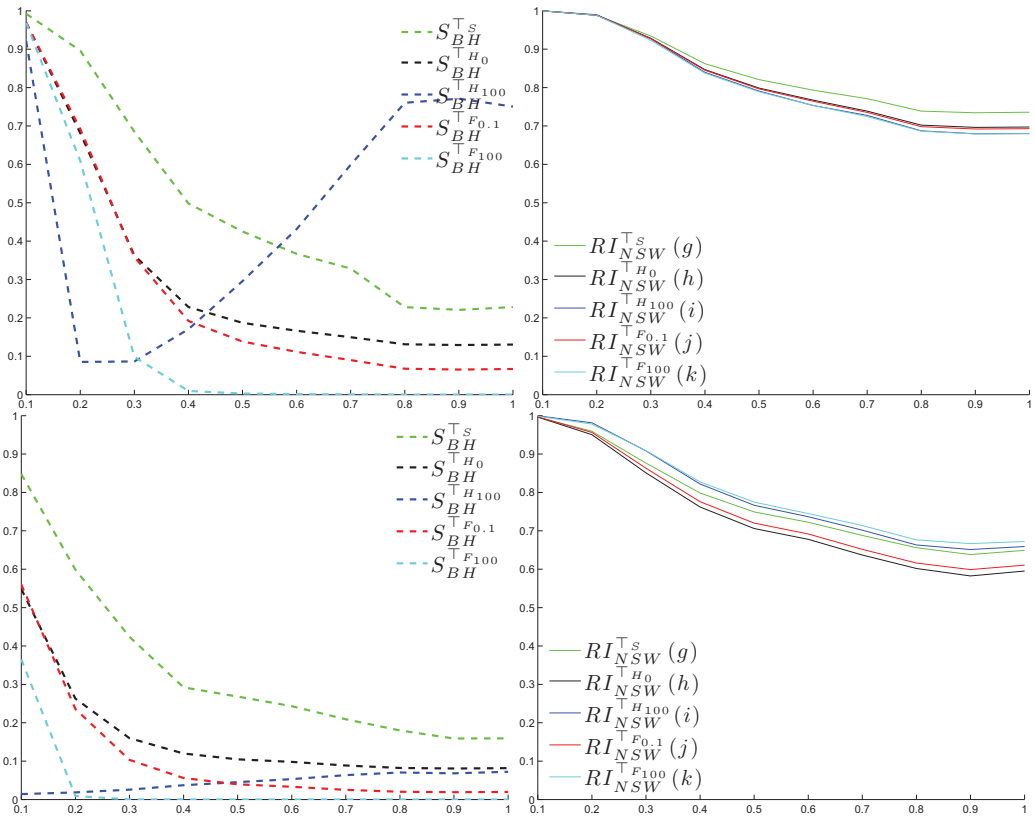



FIGURE B.22 – Comparaison de deux collections de partitions non strictes à une partition stricte de référence $U_{k^\star}$. Première ligne : partitions floues $V_\sigma$. Deuxième ligne : partitions possibilistes $V'_\sigma$.

### B.2.3 Partitions non-strictes de données réelles (E4)

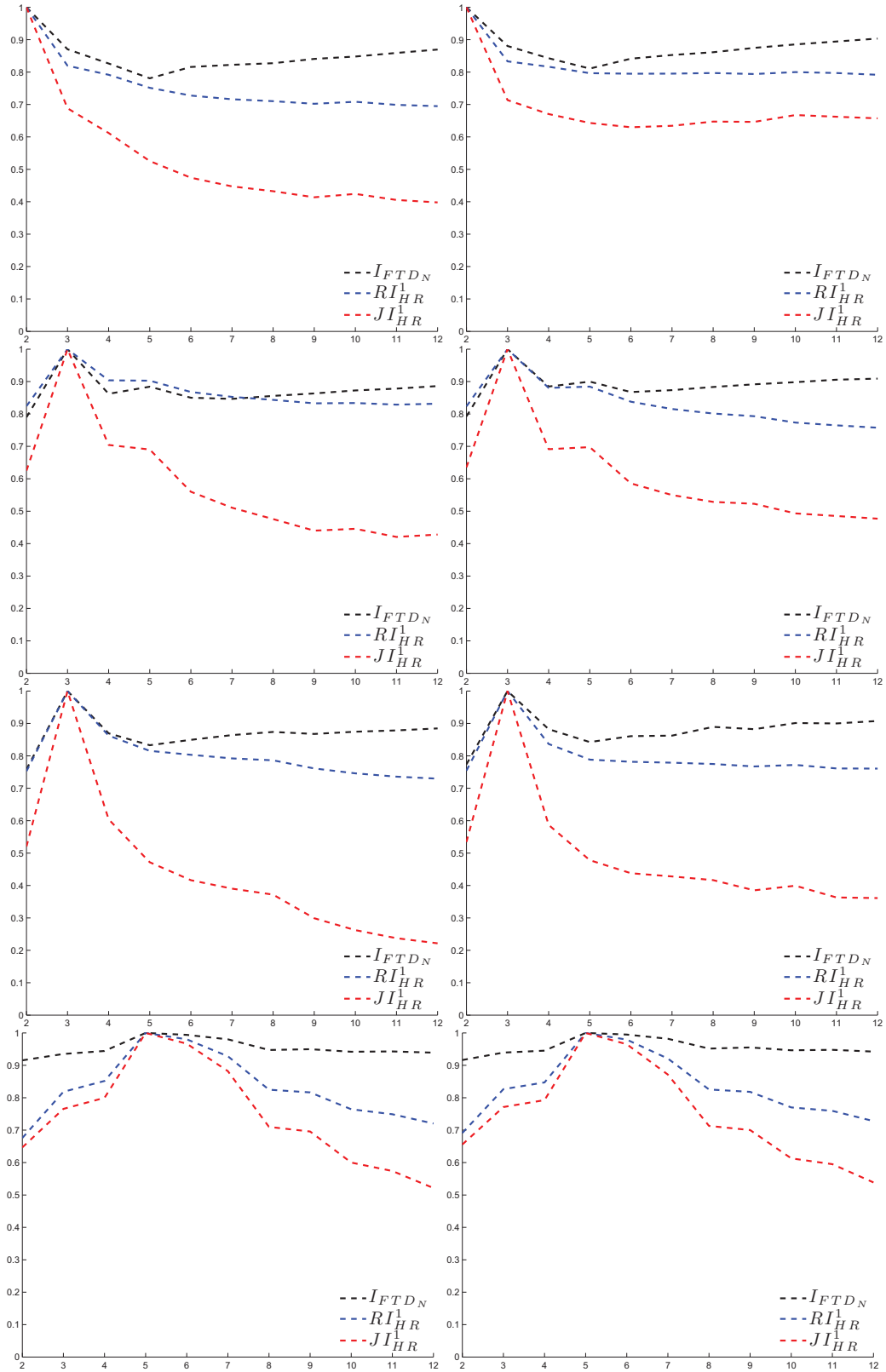



FIGURE B.23 – Valeurs de $I_{FTD_N}$, $RI^1_{HR}$ et $JI^1_{HR}$ pour la comparaison de partitions de $\mathbb{M}_{scn}$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks. À gauche : partitions floues de $S$. À gauche : partitions possibilistes de $S'$.

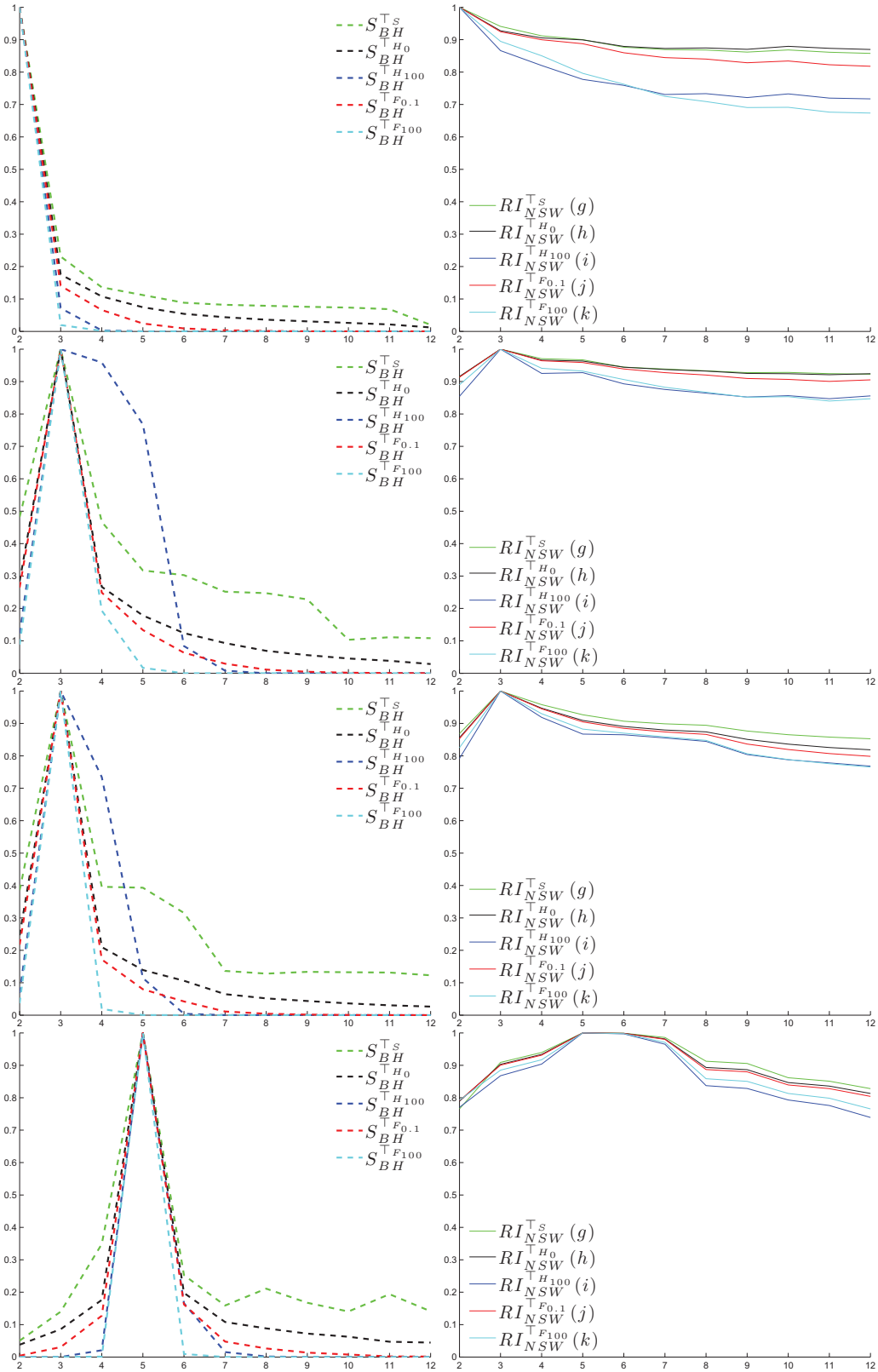



FIGURE B.24 – Valeurs de $S_{BH}^{\top}$ et $RI_{NSW}^{\top}$ pour la comparaison de partitions floues $U_c$ avec une partition floue de référence $R_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

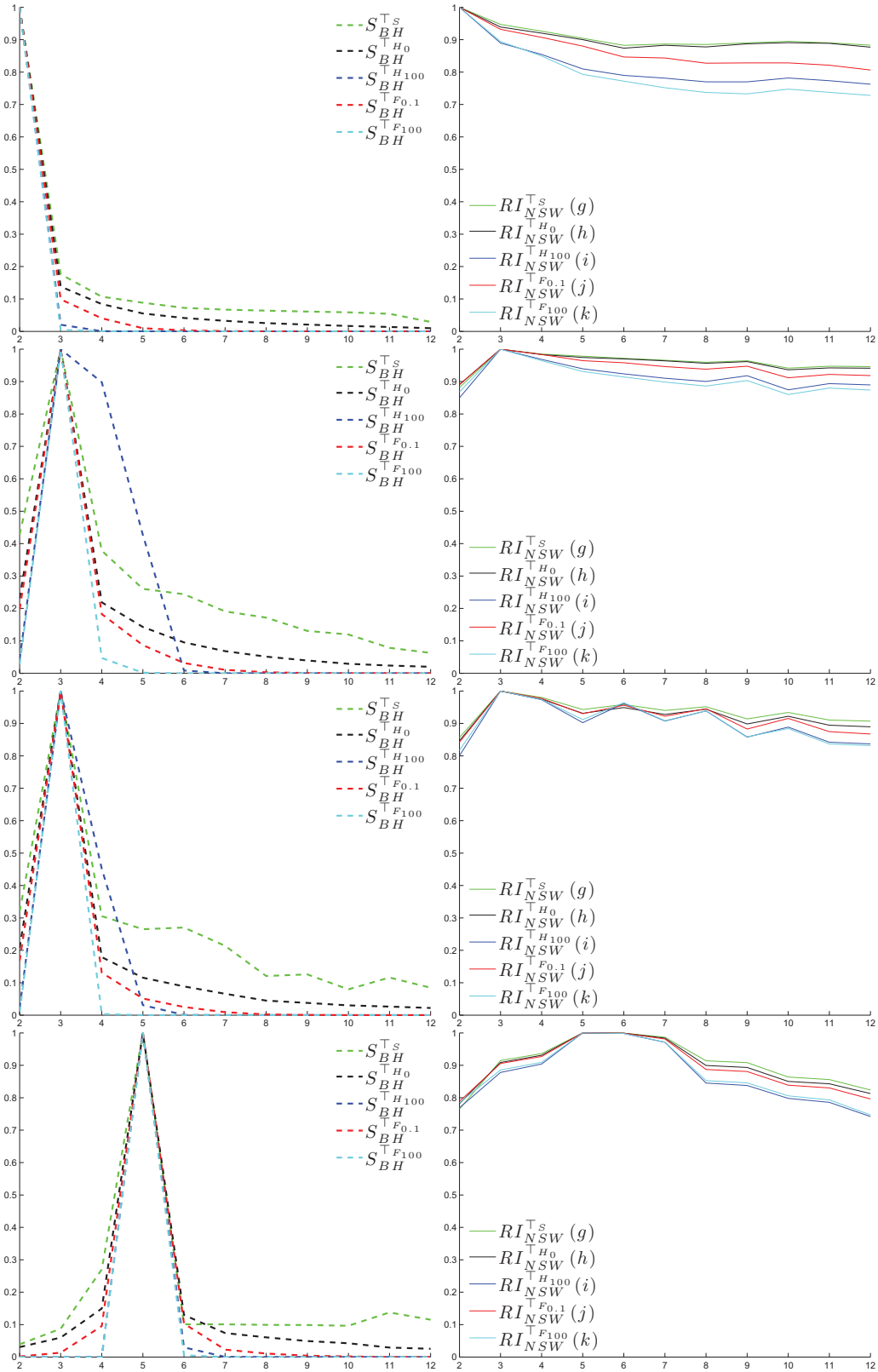



FIGURE B.25 – Valeurs de $S_{BH}^{\top}$ et $RI_{NSW}^{\top}$ pour la comparaison de partitions possibilistes $U'_c$ avec une partition possibiliste de référence $R'_{c^\star}$ pour $c = 2...12$. De haut en bas : données Pima, Iris, Wine et Pageblocks.

# Table des figures

# Liste des tableaux