THESE DE DOCTORAT DE L'UNIVERSITE PARIS 6

Spécialité

MATHEMATIQUES (STATISTIQUE)

Présentée par

Genane YOUNESS

Pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITE PARIS 6

Sujet de la thèse :

Contributions à une méthodologie de comparaison de partitions

Date de soutenance : 1 juillet 2004

Devant le jury composé de :

MM.	Israël- César LERMAN	Président
	Gilbert SAPORTA	Directeur de thèse
	Gilles CELEUX	Rapporteur
	Jean- Paul RASSON	Rapporteur
	Paul DEHEUVELS	Examinateur
	Pierre CAZES	Examinateur

Remerciements

Je tiens à remercier vivement le professeur Gilbert SAPORTA, chaire de la statistique appliquée au C.N.A.M- Paris, pour la confiance qu'il m'a témoignée en acceptant la direction scientifique de mes travaux. Je lui suis reconnaissante de m'avoir fait bénéficier tout au long de ce travail de sa grande compétence, de sa rigueur intellectuelle, de son dynamisme, et de son efficacité certaine que je n'oublierai jamais. Soyez assuré de mon attachement et de ma profonde gratitude.

Je suis très honoré à remercier de la présence à mon jury de thèse et je tiens à remercier :

Monsieur Paul Deheuvels, directeur du laboratoire LSTA, pour l'honneur qu'il m'a fait en acceptant d'être membre de mon jury de thèse. Je tiens à l'assurer de ma profonde reconnaissance pour l'intérêt qu'il porte à ce travail.

Monsieur Gilles CELEUX, directeur de recherche à l'INRIA, pour l'honneur qu'il m'a fait pour sa participation à mon jury de thèse en qualité de rapporteur de mon travail, pour le temps consacré à la lecture de cette thèse, et pour les suggestions et les remarques judicieuses qu'il m'a indiquées.

Monsieur Jean Paul RASSON, professeur aux Facultés universitaires Notre Dame de la Paix à Namur, pour sa participation à mon jury de thèse en qualité de rapporteur de mon travail et pour toutes remarques intéressantes qu'il m'a faites.

Monsieur Pierre CAZES, professeur à l'université Paris-9 Dauphine, d'avoir accepté de faire parti du jury de cette thèse. Je le remercie pour les conseils scientifiques

qu'il a apporté en qualité d'éditeur de la revue de la statistique appliquée, en jugeant une partie de cette thèse, ainsi que pour son immense aide pour mener à bien ces travaux.

Monsieur Israël- César LERMAN, professeur à l'université de Rennes, qui a bien voulu juger une grande partie de ce travail en tant que rapporteur du journal STUDENT. Je le remercie pour le temps consacré à la lecture de ce travail ainsi que pour les commentaires m'ayant permis de l'améliorer.

Monsieur Yves Lechevallier, directeur de recherches à l'INRIA, pour l'intérêt qu'il a manifesté en participant en qualité de membre invité à ce jury.

Je tiens également à exprimer ma reconnaissance à Monsieur Youssef ABOU NADER, ancien directeur de l'Institut des Sciences Appliquées et Economiques, centre du Liban associé au CNAM- Paris, pour son soutien permanent aussi bien pour mes travaux de recherches que pour mes enseignements.

Je remercie Monsieur Hassan AWADA, pour son intérêt permanent à mon égard et pour son soutien sur le plan humain.

A titre plus personnel, Je remercie chaleureusement mon mari, Bilal, pour la grande patience, l'encouragement et la confiance qu'il m'a témoigné dont il a fait preuve à la relecture de mon manuscrit. Je tiens à le remercier surtout pour son soutien moral ininterrompu et ses nombreux conseils tout le long de ma thèse.

Finalement je remercie mes parents pour leurs soutiens qui m'a été bien utile durant ma thèse.

Résumé

La comparaison de classification est l'une des questions ouvertes en analyse de données. Le besoin de comparer deux partitions survient lors de l'étude de deux enquêtes portant soit sur les mêmes individus, soit sur un même questionnaire. L'objectif de notre travail est d'étudier ces différentes approches et de trouver des procédures formalisées qui reposent sur des modèles probabilistes d'écart à une typologie qui soient réalistes pour le cas de comparaison de deux partitions dans les différents contextes.

Dans notre thèse, nous proposons une procédure pour comparer deux partitions proches. Notre approche consister à étudier la distribution de divers indices d'associations en engendrant par simulation des partitions qui devraient être proches car issues d'un même modèle sous-jacent qui est le modèles des classes latentes. Nous présentons les écritures contingentielles et relationnelles de ces indices de comparaison et nous cherchons leurs distributions d'échantillonnage sous l'hypothèse de liaison forte.

Pour comparer des partitions, basées sur les mêmes variables, nous proposons une méthode par projection de partitions utilisant l'analyse discriminante linéaire sur l'une des partitions et le reclassement des individus de l'autre partition sur les classes de la première. Nous présentons une autre approche basée sur l'utilisation de la classification des variables qui consiste en particulier à comparer les arbres hiérarchiques à partir d'indices de consensus.

Mots- clés: classes latentes, partition, indices d'associations, analyse discriminante linéaire, classification des variables, indices de consensus.

Abstract

Comparing partitions is one of the open-ended questions in data analysis. The need to compare two partitions occurs during the study of two surveys having the same data set or the same questionnaires. The goal of our work is to study these different approaches and to find formal procedures based on probabilistic models that are realistic in the case of comparing close partitions.

In our theses, we propose a methodology to compare two "near-identical partitions". Our approach consists in studying the empirical distribution of some association measures by simulating similar partitions coming from a common latent class model. We present the contingential and the paired comparisons forms for the association measures. We study the empirical distribution for these indexes under the hypothesis of close partitions.

For comparing partitions of different units based on the same questionnaires, we propose a method of projection of partitions using linear discriminant analysis on one of the partitions and allocating the units of the other partition in the classes of the first one. We present another approach based on the use of the classification of variables for which the procedure consists in comparing these classification according to consensus indices.

Keywords. Latent class, partitions, association indices, linear discriminant analysis, classification of variables, consensus indices.

TABLES DES MATIERES

Introduction Générale

L'une des questions ouvertes en classification est la comparaison des structures de données. Le besoin de comparer des partitions obtenues par plusieurs méthodes de classification ou sur différentes données survient lors de l'étude de deux enquêtes portant soit sur les mêmes individus, soit sur différents échantillons pour un même questionnaire.

Plusieurs auteurs se sont intéressés au problème de comparaison de partitions :

Rand, W.M. [RAN 71] a proposé l'indice d'accord considéré comme le mieux adapté à cette problématique. Cet indice a été ensuite modifié par Fowlkes, E.B. et Mallows, C.L. [FOW 83]. Basé sur la comparaison des triples objets, Hubert L., et Arabie, P. [HUB 85] ont proposé l'utilisation de cet indice pour mesurer la correspondance entre les partitions. En utilisant l'aspect mathématique et statistique des coefficients de comparaison, Lerman, I.C. [LER 88] a tenu compte des contraintes relationnelles qui résulte de la structure d'une partition. Une présentation de l'indice de Rand en utilisant le concept de comparaison par paires, a été réalisée par Marcotorchino, J.P. [MAR 91]. En 1997, Saporta, G. [SAP 97] a présenté diverses approches destinées à répondre aux questions suivantes lors de la comparaison de deux enquêtes: « peut-on affirmer que la classification n'a pas changé, que le nombre de classes est le même, que les proportions respectives des classes ont ou n'ont pas varié, que les classes s'interprètent de la même façon ? ». Une méthode de recherche d'une classification consensus à partir de plusieurs partitions, utilisant l'indice de Rand a été proposée par Krieger, A. et Green, P. [KRI 99].

Une fois définie un indice de similarité entre partitions, une manière d'aborder le problème de la comparaison consiste à calculer une valeur critique au-dessus ou en deçà de laquelle on conclura que les deux partitions sont ou non concordantes.

Il faut alors connaître la distribution de probabilité de cet indice, mais sous quelle hypothèse ? Cette question ne semble curieusement pas avoir été traitée dans la

littérature, en tous cas pas sous des hypothèses réalistes [SAP 97, 01, 02]. En effet, les rares travaux connus et récents [IDR 00], concernent la distribution de l'indice de Rand et de l'indice de Janson et Vegelius sous l'hypothèse d'indépendance. Or cette hypothèse n'est évidemment pas pertinente pour la question posée, car la non-indépendance ne signifie pas une forte concordance. La difficulté est de conceptualiser une hypothèse nulle d'identité de deux partitions. Nous nous trouvons dans une situation voisine de celle où nous voudrons tester que deux variables numériques sont identiques : or si $\rho=1$, nous savons que $r=1$ et nous n'avons pas de test utile de l'hypothèse nulle qui se trouve rejetée dès que $r>1$.

L'objectif de notre travail est d'étudier ces différentes approches et de trouver des procédures formalisées qui reposent sur des modèles probabilistes d'écart à une typologie qui soient réalistes pour le cas de comparaison de deux partitions sur le même ensemble d'individus ou sur un même groupe de variables. Ces procédures tiennent comptes du fait que l'appartenance à une classe comporte toujours une part d'incertitude.

Objectifs et Originalité de la thèse

Nos recherches sont axées sur le problème de comparaison de classifications en analyse de données. Notre objectif sera dans un premier temps de trouver une méthodologie pour comparer des partitions provenant d'un même ensemble de données. Nous présentons les écritures relationnelles et contingentielles des différents indices de concordance et nous cherchons leurs distributions d'échantillonnage sous l'hypothèse d'absence de liaison.

Pour définir ce que nous entendons par « partitions proches », notre approche consiste à dire que les individus proviennent d'une même partition commune, dont les deux partitions observées en sont des réalisations bruitées. Nous construisons à partir d'une partition initiale basée sur des caractéristiques probabilistes (le modèle des classes latentes), deux partitions par la méthode des k-means. Ces deux partitions qui ne diffèrent que d'une façon aléatoire sont comparées à partir des indices de ressemblance. Une étude distributionnelle de ces différents indices est effectuée.

Nous proposons une nouvelle méthode de comparaison de partitions, basée sur les mêmes variables, par projection de partitions. Notre procédure consiste à appliquer l'analyse discriminante sur une des deux partitions et à reclasser les individus de l'autre partition

sur la première partition. Toujours dans le même contexte de comparaison, nous donnons une autre approche basée sur l'utilisation de la classification des variables dont la démarche est de trouver les arbres hiérarchiques et de les comparer à partir des indices de consensus.

Plan de la thèse

Dans un premier chapitre, nous évoquons brièvement quelques travaux réalisés concernant les problèmes de l'existence, de la détermination du « vrai » nombre des classes d'une partition ainsi que les algorithmes de classification. Les modèles probabilistes qui évaluent et étudient l'existence d'une partition sont évoqués.

Dans un deuxième chapitre, on présente les méthodes classiques utilisées en analyse de données et dans le cadre de l'analyse des données symboliques, on s'intéresse aux travaux offrant une aide à l'interprétation des résultats, au moyen de règles logiques, la méthode « CABRO », le marquage sémantique, et la méthode de classification divisive.

Le troisième chapitre étudie en détail les différents indices qui serviront par la suite pour notre étude : Rand, Rand asymétrique, kappa de Cohen etc. Des formulations contingentielles et relationnelles pour la plupart de ces indices sont présentées.

Au quatrième chapitre, nous cherchons à comparer deux partitions provenant d'un même ensemble d'individus décrits par deux ensembles de variables pour tester si elles sont proches ou non. Nous nous intéressons à tester la stabilité des classes et de leurs interprétations pour les deux partitions. Nous présentons ici une méthodologie de constructions de partitions proches utilisant un modèle de classes latentes.

Le cinquième chapitre est consacré à la présentation des tests classiques de comparaison des deux échantillons. Nous proposons une nouvelle méthode de comparaisons par projection des partitions. Une autre approche pour la comparaison de partitions dans notre cas est définie par utilisation de la classification des variables. Enfin, la stabilité des interprétations des classes des partitions étudiées est traitée.

Dans le but de valider l'étude présentée dans les deux derniers chapitres, le dernier chapitre est consacré à l'application des différents algorithmes sur des données réelles.

Une partie des résultats des chapitres trois et quatre a fait l'objet de publications :

Genane Youness, Gilbert Saporta: Some Measures of Agreement Between Close Partitions- *Journal Student,* à paraître en 2004.

Genane Youness, Gilbert Saporta: Une Méthodologie pour la Comparaison de Partitions- *Revue de la Statistique Appliquée*, vol. LII (1), 97-120, 2004.

Genane Youness, Gilbert Saporta: Sur des Indices de Comparaison de deux Classifications - In *Proceedings SFC03, 10 èmes rencontres de la Société Francophone de Classification*, 177-180, Neuchâtel 10-12 septembre 2003.

Gilbert Saporta, Genane Youness: Comparing Two Partitions: Some Proposals and Experiments - *In Proceedings in Computational Statistics 2002,* 15th Symposium held in Berlin, Germany, 2002 Haerdle, Wolfgang; Roenz, Bernd (Eds.), Physica Verlag, 243-248, 2002.

Gilbert Saporta, Genane Youness: Concordance entre Deux Partitions: Quelques Propositions et Expériences - *In Proceedings SFC 2001. 8èmes rencontres de la Société Francophone de Classification*, Pointe à Pitre, décembre 2001.

Chapitre 1

Panorama sur quelques Méthodes et Problèmes de Classifications

1.1 Introduction

Dès les premières tentatives de classification s'est posé le problème du nombre de classes, de la validation, et de l'existence de ces classes. Les questions sont aussi simples que les réponses sont complexes : Existe t- il des classes ? Si oui, combien ? Une fois trouvées des réponses, on peut donc traiter la comparaison de deux partitions provenant d'un même ensemble de données ou d'un même questionnaire.

Nous évoquerons brièvement dans ce chapitre quelques travaux réalisés à propos des problèmes de l'existence, de la détermination du « vrai » nombre des classes d'une partition ainsi que les algorithmes de classification.

Nous présenterons en premier lieu les modèles probabilistes qui évaluent et étudient l'existence d'une partition, parmi lesquels les modèles de partitions fixes, et les modèles de classes latentes que nous utiliseront par la suite dans notre travail pour générer des partitions proches. La deuxième partie présente les algorithmes de classification les plus utilisés comme la méthode de classification autour de centres mobiles, les k-means et les algorithmes ascendants. La troisième partie est consacrée aux approches de validation qui se trouvent dans la littérature dans un cadre non probabiliste. Les tests statistiques de F maximum, de Wilks maximum et de gap peuvent être utilisés pour tester l'homogénéité des classes des partitions. En dernier lieu, les critères AIC, BIC, MLD, ICOMP, les critères d'entropie EC et NEC seront présentés comme des méthodes visant à déterminer le nombre des classes d'une partition.

1.2 Modèles probabilistes

L'approche probabiliste de la classification consiste à supposer que les individus à classer soient des réalisations indépendantes d'une variable aléatoire de distribution de probabilité F, et à baser la recherche d'une classification sur l'analyse de cette loi F. Les hypothèses que l'on est amené à poser pour que F induise un partitionnement de la population définissent différents modèles probabilistes de classification : modèle de multimodalité, le modèle de partitions fixes et le modèle de mélange. Pour notre travail, c'est ce dernier que nous choisissons pour générer des partitions à bases probabilistes.

1.2.1 Modèles de partitions fixes

Ce modèle a été proposé par Scott et Symons en 1971, il suppose l'existence d'une partition inconnue à k classes d'effectifs respectifs n_1, n_2,…, n_k et tel que $\sum_{h=1}^{k} n_h = n$. A chacune des k classes est associée une densité $f_h(x)$. Si les $f_h(x)$ sont des lois normales sphériques de même matrice de covariances $\sigma^2 I$ et de moyennes μ_h, la partition qui réalise le maximum de vraisemblance est celle qui minimise le critère suivant [LEB 97]:

$$cr(k) = \sum_{h=1}^{k} \sum_{i \in P_h} \left\| x_i - \bar{x}_h \right\|^2 = \sum_{h=1}^{k} \sum_{i \in P_h} d^2(xi, g_h)$$

où g_h est le centre de gravité de la classe h de composantes $\bar{x}_h$.

On retrouve dans ce cas particulier le critère d'inertie utilisé dans la méthode de Forgy ou des nuées dynamiques, ce qui permet de comprendre pourquoi ces méthodes ont tendance à créer des classes sphériques.

1.2.2 Modèles de mélanges

Le modèle de mélanges de distribution est le modèle théorique de base le plus répandu en classification. Plusieurs travaux ont été faits sur l'estimation des mélanges de densité [BOC 77], [CEL 92], etc. L'observation x_i (i<=n) est une réalisation d'une variable aléatoire x de densité f(x) :

$$f(x) = \sum_{h=1}^{k} p_h f_h(x) \text{ avec } 0 < p_h < 1 \text{ et } \sum_{h=1}^{k} p_h = 1$$

$f_h(x)$ est la densité de la classe h dont la forme doit être spécifiée. Le nombre de classe k doit être connu. Dans ces conditions, l'hypothèse d'absence de structure peut être celle de l'identité de diverses composantes $f_h(x)$ de la densité f(x).

1.3 Classes latentes

Introduit par Lazarsfeld, P.F[LAZ 50], le modèle des variables latentes est un modèle de mélange qui fournit un important outil pour l'analyse de données multivariées. Il postule l'existence de variables inobservables mais dont on peut mesurer ou observer des conséquences ou des effets. L'hypothèse fondamentale est que les covariations entre variables observées (dites également « variables manifestes ») s'expliquent par la dépendance de chaque variable observée avec les variables latentes.

Connaître les variables latentes permettrait donc de diminuer les corrélations entre variables observées, d'où le principe fondamental d'indépendance conditionnelle : les variables observées sont indépendantes conditionnellement aux variables latentes. L'analyse en facteurs communs et spécifiques en est le cas particulier le plus connu, où variables observables et facteurs sont tout quantitatifs. Le tableau suivant présente les différentes situations selon la terminologie de Bartholomew et Knott [BAR 99] :

	Variables latentes	
Variables observées	**qualitatives**	**quantitatives**
qualitatives	Analyse des classes latentes	Analyse des traits latents
quantitatives	Analyse des profils latents	Analyse factorielle

Tab. 1.1 *Classification des méthodes des variables latentes*

Les modèles des classes latentes et les modèles de profils latents sont à la classification métriques, ce que l'analyse factorielle est à l'analyse en composantes principales.

1.3.1 Les classes latentes

Définition

Le modèle des classes latentes caractérise souvent des variables discrètes multidimensionnelles latentes à partir de variables binaires observées; On peut illustrer la forme générale de l'analyse des classes latentes examinée par [LAZ 68], et [GOO 79].

On considère un ensemble de p variables observées dichotomiques X_1, X_2,..,X_p prenant des valeurs 0 ou 1, et Y la variable latente à k classes, on notera p_{jh} la probabilité que X_j=1 pour un individu de la classe latente h. Si π_h est la probabilité *a priori* d'appartenir à la classe latente h, l'hypothèse d'indépendance conditionnelle donne pour le vecteur des variables observées x:

$$f(x) = \sum_{h=1}^{k} \pi_h \prod_{j=1}^{p} p_{jh}^{x_j} (1 - p_{jh})^{1-x_j}$$

On en déduit que la probabilité *a posteriori* d'un individu de vecteur x appartenant à la classe latente h est :

$$H(h/x) = \pi_h \prod_{j=1}^{p} p_{jh}^{x_j} (1 - p_{jh})^{1-x_j} / f(x)$$

La formule précédente permet donc d'affecter un individu à la classe latente la plus probable. Le problème statistique se ramène donc à estimer les paramètres π_h et p_{jh}, et à tester l'ajustement du modèle.

Le modèle de classes latentes peut s'étendre sans difficultés à des variables observées à plus de deux catégories, mais ne sera pas développé ici.

L'estimation des paramètres

Prenons un échantillon de n observations et notons x_{ji} la valeur prise par X_j pour l'individu i. On utilise la méthode du maximum de vraisemblance, la log-vraisemblance vaut :

$$l = \sum_{i=1}^{n} \ln\left(\sum_{h=1}^{k} \pi_h \prod_{j=1}^{p} p_{jh}^{x_{ji}} (1 - p_{jh})^{1-x_{ji}} \right)$$

Depuis Goodman [GOO 79], la maximisation de *l* s'effectue à l'aide de la méthode itérative EM, qui semble la mieux adaptée[BAR 99]. Comme $\sum_{h=1}^{k}\pi_h = 1$, on maximise le lagrangien $\phi = l + \lambda \sum_{h=1}^{k}\pi_h$ en simplifiant avec la formule de Bayes et en introduisant les probabilités *a posteriori*, on trouve que $\lambda = -n$.

La première équation d'estimation est:

$$\hat{\pi}_h = \sum_{i=1}^{n} H(h/x_i)/n$$

La deuxième :

$$\sum_{i=1}^{n} (x_{ji} - p_{jh}) H(x_i/h)/p_{jh}(1-p_{jh}) = 0$$

soit

$$\hat{p}_{jh} = \sum_{i=1}^{n} x_{ji} H(h/x_i)/n\hat{\pi}_h$$

On peut, grâce aux équations de la vraisemblance, obtenir des erreurs standards asymptotiques, mais divers auteurs conseillent plutôt d'utiliser le bootstrap, en particulier si n est faible.

Ajustement et choix de modèles

Une fois les paramètres estimés, on peut alors comparer les fréquences observées n(x) des différents vecteurs x possibles (2^p au maximum) de variables observées, avec leurs espérances données par $n\hat{f}(x)$.

On compare alors $G^2 = 2\sum_x n(x)\ln(\frac{n(x)}{n\hat{f}(x)})$ à un khi-deux à $\nu = 2^p - k(p+1) + 1$ degré de liberté si toutes les combinaisons de réponses ont été observées avec un effectif suffisant. Il y a en effet k – 1 probabilités π_h , et kp probabilités conditionnelles p_{jh} à estimer, soit k(p+1) – 1 paramètres. Le modèle d'indépendance conditionnelle à k classes latentes est acceptable si G^2 est inférieur à un seuil.

En analyse exploratoire, on utilisera la même statistique pour choisir le nombre de classes : un usage courant consiste à tester des modèles emboîtés à 2, 3, 4 classes etc., et à s'arrêter dès que l'on trouve une valeur acceptable, car en général plus le nombre de classes est grand, plus le modèle s'ajuste bien.

Le problème du choix de modèle est l'un des recherches essentielles dans les thèmes de classification à l'aide des classes latentes. Actuellement, deux problèmes se posent, le premier concerne le choix du nombre de classes, le second concerne la forme de modèle qui donne le nombre de classes.

L'hypothèse, sous condition du nombre de classes, peut être testée en utilisant le test du rapport du maximum de vraisemblance standard entre un modèle de matrice de covariances limité et un autre dont la matrice de covariances est non limitée. Les tests de Wald et du multiplicateur de Lagrange peuvent être utilisés pour estimer la signification de certains termes inclus ou exclus respectivement. C'est bien connu, que le test de Khi-deux ne peut pas être utilisé pour déterminer le nombre de classes. On va voir dans la suite les plus importants critères de choix de modèle.

Les modèles de classes latentes peuvent servir dans une optique exploratoire ou confirmatoire, mais souffrent des critiques adressées classiquement à l'analyse factorielle vis à vis des méthodes de type ACP : Problèmes d'identification, d'existence des variables latentes qui ne sont jamais que des constructions, ainsi que de la non-convergence des algorithmes dans certains cas, ou de la convergence vers des extremum locaux.

1.3.2 Les profils latents

Le modèle de profils latents [VER 02, 03] est un modèle à variables latentes qualitatives et des indicateurs ou variables observées continues. Introduit par Lazarsfeld et Henry [LAZ 68] en 1968, il peut être considéré comme un modèle probabiliste pour l'analyse de classification non hiérarchique comme les méthodes des k-means. On l'appelle aussi modèle de mélange de composantes normales, modèle de mélange ou analyse discriminante latente.

Comme en analyse des classes latentes, les profils latents supposent que la population est formée de k classes ou groupes non observés qui peuvent s'appliquer aux profils latents. Les variables observées sont supposées normalement distribuées. Généralement on a des distributions normales multivariées. La densité de mélange des densités des classes latentes est donnée par la formule suivante :

$$f(x) = \sum p(h) f(x / \mu_h, \Sigma_h)$$

Où chaque classe latente h a son propre moyenne μ_h et sa matrice de covariances Σ_h. La proportion des individus dans chacune des composantes est notée p(h).

Le modèle est proche de l'analyse discriminante quadratique avec la différence que les classes ne sont pas connues. Les restrictions sur l'égalité des matrices de covariances et de leur diagonalisation sont similaires aux cas de l'analyse discriminante linéaire. Le modèle peut être écrit de la façon suivante:

$$f(x) = \sum_{h=1}^{k} p(h) \prod_{j=1}^{p} f(x_j / \mu_{jh}, \sigma^2_{jh})$$

La deuxième hypothèse d'indépendance locale et d'égalité des variances d'erreur $\sigma^2_{jh} = \sigma^2_j$ donne des spécifications proches de la méthode des k-means.

Plusieurs méthodes ont été proposées pour la structure des matrices de covariances des classes. L'utilisation des matrices diagonales par blocs est un compromis entre une matrice de covariances complète et une autre diagonale.

Divers logiciels d'estimations des profils latents ont été proposés tels que EMMIX, Mclus, Mplus et latentGOLD.

1.3.3 Utilisation du modèle de profils latents pour simuler des partitions

Plusieurs auteurs se sont intéressés à l'application du modèle de classes latentes comme une méthode de classification ou un modèle particulier de mélange de distributions. Citons [McL 88], [McC 87], [EVE 93], [CHE 95] et récemment [BAR 99].

Le modèle de classes latentes est bien adapté pour engendrer des partitions. Notons que ce modèle a été récemment utilisé pour la recherche de partitions consensus par Green et Krieger [GRE 99]. Plus précisément, comme nous allons utiliser des variables observées quantitatives, selon la terminologie de Bartholomew et Knott [BAR 99], on utilise le modèle de profits latents. L'hypothèse de base est l'indépendance des variables observées conditionnellement aux classes latentes. On sait que ce modèle souffre de problèmes sérieux d'identifiabilité, mais ici il n'est utilisé que pour engendrer des données et non pour estimer des paramètres. Il suffit alors de générer des distributions indépendantes dans chaque classe, après avoir tiré le numéro de classe de chaque observation selon une multinomiale de probabilités π_h.

1.4 Algorithmes de classifications

Il existe plusieurs familles d'algorithmes de classifications : les algorithmes conduisant directement à des partitions comme la méthode de classification autour de centres mobiles (cas particulier de techniques de nuées dynamiques ou des k-means), les algorithmes ascendants (ou agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux fournissant une hiérarchie de partitions des objets, et les algorithmes descendants (divisifs) qui procèdent par dichotomies successives des objets. On se limite dans la suite aux deux premières techniques de classifications.

1.4.1 Méthodes des nuées dynamiques

L'algorithme connu sous le nom de nuées dynamiques étudié formellement par Diday [DID 71] permet de traiter des ensembles d'effectifs assez important en optimisant localement un critère de type inertie.

Etant donnée une partition à k groupes de n points, de $g_1,..,g_k$ centres de gravités et de $I_1,\ldots,I_k$ inerties. L'inertie totale I des n points autour du centre de gravité globale g est :

$$I=I_B+I_W$$

avec $I_B=\sum n_{P_h} d^2(g_h,g)$ est l'inertie interclasse des k centres de gravités étant donnée

n_{P_h} le poids de la $h^{ième}$ classe,

et $I_W = \sum n_{P_h} I(P_h)$ est l'inertie intraclasse

Un critère usuel consiste à chercher la partition telle que I_W soit minimale pour avoir des classes homogènes pour k fixé. Ce qui revient à chercher le maximum de I_B. .

La méthode de centres mobiles peut être imputée principalement à Forgy [FOR 65], c'est un cas particulier de la méthode des nuées dynamiques. Elle consiste à déterminer k centres provisoires de classes. Ces k centres définissent une partition en k classes. Ainsi l'individu i appartient à la classe P_h s'il est plus proche de g_h que de tous les autres centres. On remplace alors les k centres provisoires par les k centres de gravités de ces classes et on recommence. L'algorithme converge vers un optimum, souvent local, qui minimise l'inertie intraclasse. Ce minimum dépend du système initial de centres en un nombre fini d'itérations.

Dans la technique des nuées dynamiques, les classes peuvent ne pas être caractérisées par un centre de gravité, mais par un noyau ayant un meilleur pouvoir descriptif que des centres ponctuels.

La méthode dite des k-means (k-moyennes) introduite par MacQueen [MAC 67] commence par un tirage pseudo-aléatoire de centres ponctuels. Chaque réaffectation d'individus entraîne une modification de la position du centre correspondant, on peut en une seule itération trouver une partition de bonne qualité mais dépendant de l'ordre des individus.

1.4.2 Classification hiérarchique ascendante

L'algorithme est du à Sokal [SOK 63], puis étudié par Lance et Williams [LAN 67], et Gordon [GOR 87]. Son principe consiste à fournir un ensemble de partitions en classes de moins en moins fines obtenues par groupements successifs de parties. La classification hiérarchique se représente par un dendrogramme ou arbre de classification.

Une famille H des parties de l'ensemble des objets Ω est une hiérarchie si :

- Ω et les singletons appartiennent à H
- $\forall A, B \in H \quad A \cap B \in \{A, B, \phi\}$

Une hiérarchie indicée est un couple (H, f) où H est une hiérarchie et f une application de H dans IR^+ telle que :

- f(A)=0 si et seulement si A ne contient qu'un seul individu.
- $\forall A, \text{et } B \text{ dans } H, A \subseteq B \text{ et } A \neq B \Rightarrow f(A) < f(B)$

Cet indice permet de valuer l'arbre hiérarchique associé et de définir un indice de dissimilarité d_H qui est défini sur Ω par :

$$d_H(x,y) = \min\{f(A)/x,y \in A, A \in H\} \ \forall x,y \in H$$

d_H est appelé dissimilarité induite par (H, f) ou ultramétrique induite par l'hiérarchie. Plus les individus se regroupent du bas de l'arbre plus ils se ressemblent au sens de cet indice.

La CAH génère cet arbre de classification de manière ascendante : on regroupe les deux individus les plus proches qui forment un sommet, il ne reste plus que n-1 individus et on itère le processus jusqu'au regroupement complet de tous les individus. Un des problèmes consiste à définir une mesure de dissimilarité entre parties.

La CAH nécessite aussi la connaissance d'une mesure de ressemblance entre groupes. Cette mesure est appelée indice ou critère d'agrégation. Les différents critères d'agrégations les plus utilisés seront exposés dans le chapitre 5.

1.5 Détermination et validation du nombre de classes

A l'issue de la classification, il est nécessaire de s'assurer de la validité des classes obtenues. Ce problème a fait l'objet de nombreux travaux, citons : Bock [BOC 94], Gordon [GOR 98], Milligan [MIL 96], Jain et Dubes [JAI 87] et Bel Mufti [BEL 98]. Trois approches de validation ont été proposées pour justifier l'existence des classes. Il est à noter que la qualité d'une partition est très liée au choix de nombre de classes.

Parmi les tests les plus répandus pour la validation des classes, évoquons le test du critère de F maximal qui n'est autre que le quotient de la variance inter-classes par la variance intra-classes, le test de Maximum de Wilks qui est le quotient des déterminants des deux matrices de covariances, le « gap » test et le test de la similarité moyenne. Nous présentons ensuite les critères de choix de modèles les plus importants pour déterminer le meilleur nombre de classes d'une partition. Ces critères sont les AIC, CAIC, MDL, BIC,

ICOMP de Bozdogan [BOZ 00], de Bock [BOC 88, 97] et les critères d'entropie EC, NEC [CEL 96], [BIE 00].

1.5.1 Validation des classes

Tous les problèmes de classifications montrent que la validation d'une structure générée par une classification automatique est indispensable. On définit tout d'abord les critères de validation, nécessaires pour étudier la validation d'une classification, puis on présente les approches de validation qui se trouvent dans la littérature dans un cadre probabiliste et non probabiliste.

Critères de validation

Un critère de validation exprime la stratégie selon laquelle une structure de classification est validée. Il existe trois types de critères de validation selon que l'on dispose ou pas d'information *a priori* sur les données : critère interne, critère externe et critère relatif.

- **Le critère externe** mesure le degré avec lequel les données confirment des informations connues *a priori* sur les données [JAI 88]. Il permet aussi de comparer les résultats d'une classification automatique à une information sur la structure des données connue *a priori.*
- **Le critère interne** mesure l'écart entre la structure engendrée par un algorithme de classification et les données, en tenant compte du biais introduit par l'utilisation d'un algorithme pour obtenir la structure de classification.
- **Le critère relatif** permet de comparer deux structures de classification. Il décide quelle structure est meilleure dans le sens plus stable ou mieux appropriée pour les données. On parle de l'indice de David-Bouldin (paragraphe 1.4.4) et de la statistique de Hubert.

Validation dans un cadre non probabiliste

Trois approches de validation ont été proposées, dans un cadre non probabiliste. La première approche consiste à mesurer l'adéquation des résultats avec la dissimilarité initiale, on mesure le lien entre la structure et les données initiales. La deuxième approche

mesure la stabilité des résultats obtenus. La troisième approche mesure l'écart entre les classifications obtenues sur un échantillon d'apprentissage et sur un échantillon test.

- Validation de la valeur de l'indice mesurant l'adéquation des résultats avec la dissimilarité initiale. L'idée, pour valider cet indice, est de tester l'hypothèse nulle H_0 d'absence de structure en classes [BAI 82]. Ce type de test est appelé test de Monte Carlo par [BAR 63] et [HOP 68] : On simule des données selon H_0 et puis on calcule la valeur de l'indice qui évalue la structure de classification générée par la méthode de classification utilisée sur les données initiales. On teste si l'indice obtenu sur les données initiales est en accord avec les valeurs obtenues sur les données simulées.

- Validation mesurant la stabilité des résultats obtenus d'une classification par rapport aux différentes perturbations que les données peuvent subir. La stabilité des résultats de la classification est mesurée par l'écart entre la structure initiale et la structure obtenue sur les données bruitées ou par la variation d'un critère mesuré sur ces deux structures. Cette validation a été étudiée par [RAN 71] et [GNA 77]. [GOR 88], et [MIL 96]. Elle mesure la stabilité d'une classification en retirant un élément de l'ensemble des données et en mesurant l'influence du retrait de cet élément sur la classification.

- Validation mesurant l'écart entre les classifications obtenues sur un échantillon d'apprentissage et sur un échantillon test. Le principe est proche de la validation croisée : on divise l'échantillon de base en deux parties A et B, on applique une méthode de classification à chacun des deux échantillons, on mesure l'écart entre la partition de B générée par la méthode de classification, à celle obtenue en affectant les éléments de B à la partition de A, en utilisant une règle d'affectation. Plus cet écart est faible, plus la partition générée sur l'ensemble tout entier est valide. Cette méthode a été développée par McIntyre et Blashfield [McI 80], Smith et Dubes [SMI 80], et Breckenridge [BRE 89].

Validation d'une classe dans un cadre probabiliste

Trois principaux problèmes de validation dans le cadre probabiliste sont la classifiabilité des données, le nombre de classes, et la stabilité des résultats où il s'agit de déterminer si

les résultats sont de même natures sur d'autres échantillons issus de la même famille de loi de probabilité [BOC 85]. La plupart des tests statistiques sur le bien-fondé d'une partition s'appuient sur la loi limite (lorsque le nombre de l'échantillon tend vers l'infini) de statistiques sous certaines hypothèses de classifiabilité et de non classifiabilité.

Il y a deux approches différentes essentielles pour ce problème de validation : l'un par des outils descriptifs, graphiques et empiriques, l'autre par des tests d'hypothèse dans les statistiques inductives. On va présenter dans la suite les tests statistiques utilisés dans la littérature. Notons que Bock, H., [BOC 89, 96] a étudié les propriétés de ces tests significatifs pour différer entre l'hypothèse d'homogénéité d'une population et l'hypothèse alternative de classification ou d'hétérogénéité. Il a donné les distributions asymptotiques de ces tests sous H_0 et les puissances asymptotiques pour les hypothèses alternatives. Le test d'homogénéité [BOC 96] des classes cherche si les classes sont distribuées selon une densité uniforme ou unimodale.

1.5.2 Tests statistiques de classifications

L'un des problèmes de classification est la qualité et la pertinence des classes obtenues par classification, la comparaison des classes ainsi que la décision des structures de ces classes. Pour ce problème, plusieurs tests statistiques ont été proposés : test de F Maximum, test de Wilks Maximum, « gap » test, et test de la similarité moyenne.

Test de la statistique de F Maximum

On veut tester la pertinence d'une partition de k classes, obtenue par minimisation de l'inertie intra-classe. La pertinence minimise le critère suivant :

$$W_n(P) = \frac{1}{n} \sum_{h=1}^{k} \sum_{x_i \in P_h} d(x_i, g_h)^2$$

où g_h désigne le centre de gravité de la classe P_h pour h=1,…,k pour toutes les partitions de $\{x_1,x_2,….,x_k\}$ en k classes.

Soit g le centre de gravité de $x_1,x_2,….,x_k$. La statistique pour effectuer ce test est :

$$F_n(P) = R_n(P)\frac{n-k}{k-1}$$

avec $R_n(P) = \frac{B_n(P)}{W_n(P)}$ où $B_n(P) = \frac{1}{n}\sum_{h=1}^{k} n_h d(g_h - g)^2$

autrement dit $B_n(P)$ est l'inertie inter-classe de la partition P.

Bock [BOC 85] a fourni une approximation de la distribution sous l'hypothèse nul de Poisson de $R_n(P)$ qui permet de déterminer la valeur critique $r_n(\alpha)$ associée au seuil α.

Test de la statistique de Wilks Maximum

Pour tester l'hypothèse d'homogénéité contre une hypothèse alternative H_M qui suppose l'existence de k classes distinctes ou si la partition optimale trouvée à partir des données est plus distincte qu'une classification obtenue par des observations $X_1,\ldots,X_k$ d'un échantillon d'une distribution uniforme ou unimodale. La statistique de ce test répond à ce problème, en maximisant le quotient du déterminant des matrices de covariances.

Il faut maximiser la statistique définie par :

$$W_n = \frac{\det \sum_{h=1}^{k} n_h (g_h - g)(g_h - g)'}{\det \sum_{h} \sum_{x_{i \in P_h}} (x_i - g_h)(x_i - g_h)'}$$

Le « gap » test

Ce test proposé par Rasson et Kubushishi [RAS 94], est fondé sur des processus de Poisson qui utilise les éventuelles zones vides entre classes. Il est efficace pour reconnaître les classes isolées.

Pour tester l'hypothèse uniforme H_G dans le cas où les $x_1, x_2,\ldots, x_n$ ont une distribution uniforme, on considère la distance euclidienne minimale pour chaque j=1,….,n, représentant la distance de voisinage le plus proche U_{nj} définie par :

$$U_{nj} = \mathrm{Min}\{\|x_j - x_v\|, v \neq j / j = 1,\ldots,n\}$$

La statistique de « gap » est la suivante :

$$D_n = \mathrm{Max}\{U_{n1}, U_{n2}, \ldots, U_{nn}\}$$

Rejeter H_G si et seulement si $D_n > c$ tel que $P(D_n > c) = \alpha$. c est estimé par $c_n(\alpha)$ [HEN 82].

Test utilisant la distance ou la similarité moyenne

Ce test utilise la similarité moyenne pour les modèles de mélanges. Notons s(x, y) l'indice de similarité qui décrit la distance minimale entre x et y, pour x, y ϵ IR^p, La moyenne de la similarité $S_{jv}=s(x_j, x_v)$ est définie par :

$$T_n= \frac{1}{C_n^2} \sum_{1\le j<v} \sum_{v\le n} S_{jv}$$

Le test de la similarité moyenne rejette H_0 si $T_n<c$. s est généré par le noyau suivant :

$$\bar{q}(y) = \int k(x)K(x-y)dx$$

$$S=S_{jv}=\bar{q}(x_j - x_v) = \int k(x-x_j)K(x-x_v)dx$$

1.5.3 Critères de choix de modèles

Par la nécessité d'introduire un concept d'évaluation du choix de modèle, plusieurs travaux ont été faits pour trouver le bon choix de modèle. Nous présentons les critères les plus importants comme AIC, CAIC, MDL, BIC, ICOMP et les critères d'entropie EC, NEC, ICL et AWE pour évaluer le nombre de classes d'une partition basée sur des modèles de mélanges. Les critères AIC, MDL, et BIC sont des mesures de vraisemblances pénalisées

Dans le cas du modèle de classes latentes de variables qualitatives dépendantes, on cherche les critères AIC, BIC et CAIC à partir de la statistique de khi-deux de maximum de vraisemblance G^2 définie au paragraphe 1.3.1 au degré de liberté ν définie de la façon suivante :

$$\nu = c_h(\prod_p c_{xp} - 1) - m_k$$

Avec c_h le nombre total des différents modèles covariées dans l'ensemble de données, c_{xp} le nombre des modalités de x et m_k le nombre de paramètres dans le modèle. La valeur de khi-deux au degré de liberté ν peut être utilisée pour déterminer si le modèle ajuste bien l'échantillon.

Dans le cas des profils latents aux variables continues, les critères AIC, BIC et CAIC se calculent à partir de la statistique du log-vraisemblance ln *l*.

AIC

Le critère d'information d'Akaïké AIC [AKA 73] est une mesure d'ajustement basée sur la théorie de l'information. Pour les classes latentes, AIC est défini par :

$$AIC= G^2-\ln(n).\nu$$

Où ln(n) est le logarithme népérien de la taille de l'échantillon n.

Pour les profils latents, Sclove [SCL 87] a proposé, l'écriture suivante du critère:

$$AIC=- 2\ln(l) +2m_k$$

avec *l* est le maximum de vraisemblance, et m_k le nombre de paramètres à estimer dans le modèle.

Le modèle à valeur minimale de AIC est choisi comme étant le meilleur modèle qui ajuste les données. Basé sur le critère d'information d'Akaïké [BOZ 87], le critère consistent CAIC dans le cas des classes latentes est égal à:

$$CAIC= G^2-[\ln(n)+1].\nu$$

Dans le cas de profils latents, il est défini par :

$$CAIC= -2\ln(l) + m_k [\ln(n)+1]$$

MDL

Le critère de la longueur de description minimale MDL a été proposée par Rissanen [RIS 89]. Il est basé sur la théorie codée utilisant l'information statistique dans les données et les paramètres. Ce critère est défini par :

$$MDL= -2\ln(l)+m_k. \ln(n)$$

BIC

Le critère d'information bayésien de Schwarz pondère différemment le nombre de paramètres. Pour le cas des classes latentes à variables qualitatives, il est égal à :

$$BIC=G^2-\ln(n).\nu$$

Pour le cas des profils latents, il est défini par :

$$BIC= - 2\ln(l) + m_k \ln(n)$$

Pour comparer entre eux les modèles, afin d'obtenir un compromis entre modèle bien ajusté et modèle parcimonieux (avec peu de paramètres), le meilleur modèle est celui qui minimise AIC ou BIC.

Pour k nombre de classes et p nombre de variables, le nombre de paramètres à estimer dans le modèle, m_k dépend de la structure de la matrice de covariances entre les composantes des classes de mélanges, en effet :

- Pour le cas de covariances générales, c'est à dire pour les matrices de covariances différentes entre les classes. m_k est égal à :

$$m_k= kp+(k-1)+kp(p+1)/2$$

- Pour le cas des matrices de covariances égales entre les classes de composantes de mélanges $\Sigma_k=\Sigma$, m_k est égal à :

$$m_k= kp+(k-1)+p(p+1)/2$$

- Pour le cas des matrices de covariances égales et diagonales entre les composantes de mélanges, m_k est égal à :

$$m_k= kp+(k-1)+p$$

- Pour le cas où les variables auraient la même matrice de covariances et elles sont indépendantes entre les classes de composantes de mélanges (modèle sphérique), m_k est égal à :

$$m_k= kp+(k-1)+1$$

ICOMP

Le critère d'information complexe de Bozdogan ICOMP [BOZ 88, 2002] est développé d'une part sur le concept d'AIC et d'autre sur les concepts et les indices de la complexité d'information. La procédure est basée sur la complexité structurelle d'un élément ou d'un ensemble de vecteurs aléatoires à travers la généralisation de l'indice complexe de

covariance introduit par Van Emden [VAN 71]. Il est utilisé pour faire un bon choix de modèle pour les modèles de structures multivariées linéaires ou non linéaires. C'est une mesure performante pour choisir le nombre de classes.

Dans ICOMP, la complexité n'est pas considérée comme le nombre des paramètres à estimer comme le fait AIC mais comme le degré de l'interdépendance à travers des composantes du modèle. ICOMP est défini par :

$$ICOMP=-2\ln(l)+2.C(\Sigma_{Modèle})$$

Avec l est la fonction du maximum de vraisemblance, C représente la valeur réelle de mesure de complexité et $\Sigma_{Modèle}$ est la matrice de covariances estimée des paramètres du modèle.

Plusieurs formes de ICOMP ont été proposées dont la forme la plus générale est celle qui utilise la base d'information complexe de la matrice inverse d'information de Fisher ICOMP(IFIM) définie par :

$$ICOMP(IFIM)=-2\ln(l)+2C_1(\hat{F}^{-1})$$

Avec C_1 est la complexité d'information maximale de l'estimation $\hat{F}^{-1}$ de IFIM.

EC

Le critère d'entropie EC est proposé par [CEL 92, 96] pour évaluer le nombre de classes d'une partition fondée sur un modèle de mélange de lois de probabilités. Il se déduit d'une relation liant la vraisemblance $l(K)$ et la vraisemblance classifiante CL(K) d'un mélange.

Notons que $t_{ik}=\frac{p_k f(x_i,a_k)}{\sum_{h=1}^{k} p_h f(x_i,a_h)}$, est la probabilité conditionnelle que x_i résulte des k^{ieme} composantes de mélange ($1\le i\le n$ et $1\le k\le K$).

On a : $CL(K)=l(K)-EC(K)$

On trouve ainsi le critère d'entropie EC : $EC(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}\ln(t_{ik})$

Ce critère d'entropie mesure la capacité du modèle de mélange à k composants de fournir une partition de classes bien séparées. Une version normalisée de ce critère a été proposée par [CEL 96] et définie par :

$$NEC(K) = \frac{EC(K)}{l(K) - l(1)}$$

Une valeur minimale de NEC(K) estime le nombre de classes d'une partition résultante du modèle de mélange. Notons que EC(1)=0 et $l(1)$=C(1) est le maximum de vraisemblance pour une seule distribution Gaussienne.

[BIE 00] a proposé le critère ICL(K) de la vraisemblance complète intégrée pour évaluer un modèle de mélange dans une classification. ICL est approximé par l'utilisation du critère d'information bayésien BIC. Il peut trouver le nombre de classes d'une partition sensible et se montre plus robuste que le BIC dans le cas d'une violation du modèle de mélange.

Finalement, il faut mentionner que Banfield et Raftery [BAN 93], inspirés par BIC, ont suggéré une solution bayésienne, pour choisir le nombre de classes, basée sur des approximations des intégrations de vraisemblance classifiante. Cette approximation donne le critère suivant:

$$\text{AWE(K)=-2C(K)+2 } m_k \left(\frac{3}{2} + \ln n\right)$$

1.5.4 Détermination du nombre de classes

Un problème très lié au précédent est la détermination du bon nombre de classes. Pour obtenir le bon nombre de classes, il faut choisir un critère d'arrêt. Ces critères permettent, dans le cas d'une structure hiérarchique, d'identifier à quel niveau de la hiérarchie il faut arrêter d'agréger et obtenir la partition optimale.

L'un des critères est basé sur le principe de trouver le nombre de classes qui minimise les critères de classifications évoqués auparavant. Pour avoir le bon nombre de classe, on teste séquentiellement l'hypothèse d'homogénéité contre l'existence de 2, 3, 4, …classes. Milligan et Cooper [MIL 85] ont étudié 30 critères pour déterminer le bon choix et ils ont

trouvé que le test du F Maximum s'avère le meilleur. Parmi ces critères, on trouve l'indice de Davis- Bouldin donné par la quantité suivante :

$$D_k = \frac{1}{k}\sum_h R_h$$

avec $R_h = \max_{j \neq h} \frac{S_h + S_j}{T_{jh}}$ et $S_h^2 = \frac{1}{n_h}\sum_{i=1}^{n_h}(x_i^h, g_h)^t(x_i^h - g_h)$

n_h représente le nombre des éléments dans la classe P_h, et T_{jh} la distance euclidienne entre g_h et g_j. Le minimum de la courbe donnant l'indice D-B en fonction du nombre de classes correspond au bon nombre de classes.

Une autre méthode pour la détermination du nombre de classes d'une partition consiste à minimiser les critères du choix de modèles présentés précédemment [BOZ 94], [BRY 94].

Toujours dans la même approche, on peut utiliser la plus petite valeur propre des matrices de taux d'information du type F_M, estimé par l'information de Fisher, pour les classifications fixes et le modèle de mélanges respectivement [WIN 94].

Jain et Moreau [JAI 87] proposent un algorithme d'estimation du bon nombre de classes en se basant sur le technique du boostrap [EFR 79]. L'algorithme consiste à générer n échantillons par la technique du bootstrap, un programme de k-means est utilisé pour obtenir les partitions de chaque ensemble de données avec plusieurs nombres de classes. On calcule, pour chaque nombre de classes, le critère de la stabilité. La combinaison de ce critère avec le critère de compacité des k-classes des partitions forme la statistique qui caractérise la vraie valeur de k : la valeur de k, qui minimise cette statistique, est le nombre de classes estimé.

Halkidi [HAL 01] a proposé un indice de validation d'une classification, S-Dbw, basée sur des critères de classification, permettant de sélectionner les paramètres optimaux pour une meilleure partition. Elle utilise le critère relatif qui travaille sur la grande séparation des classes et sur la compacité maximale d'une classe de la partition. Pour une partition à c classes, v_i est le centre de la classe i, et u_{ij} est le milieu du segment $[v_i v_j]$, S-Dbw est défini par :

$$\text{S-Dbw(c)} = S_{catt}(c) + D_{ens\text{-}bw}(c)$$

Où la variance intra-classe qui indique les classes compactes, $S_{catt}(c)$ est définie par :

$$S_{catt}(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|}{\|\sigma(S)\|}$$

Avec $\sigma(S)$ est la variance de l'ensemble de données et sa p^{ieme} dimension est définie par :

$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}(x_k^p - \bar{x}^p)^2$ et $\bar{x}^p$ est la p^{ieme} dimension de $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall k \in S$

et $\sigma(v_i)$ est la variance de la classe c_i et pour la p^{ieme} dimension vaut :

$$\sigma_{v_i}^p = \frac{\sum_{k=1}^{n_i}(x_k^p - v_i^p)^2}{n_i}$$

et la densité inter-classe qui indique la séparation des classes, est définie par :

$$D_{ens\text{-}bw}(c) = \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{j=1}^{c}\frac{densite(u_{ij})}{\max\{densite(v_i), densite(v_j)\}}$$

Où la densité est définie par :

Densité(u)= $\sum_{l=1}^{n_{ij}} f(x_l, u)$ sachant que la fonction f(x, u)= $\begin{cases} 0 & \text{si } d(x,u) > stdev \\ 1 & \text{ailleurs} \end{cases}$

$$\text{Avec stdev} = \frac{1}{c}\sqrt{\sum_{i=1}^{c}\|\sigma(v_i)\|}$$

C'est évident qu'un point appartient au voisinage de u si sa distance de u est plus petite que la moyenne écart type des classes stdev.

La valeur c qui minimise l'indice de validité S-Dbw(c) peut être considérée comme étant la valeur optimale pour le nombre de classes d'une partition présentes dans l'ensemble de données en se basant sur les deux critères de compacité de la séparation des classes.

Ben- Hur et al. [BEN 02] proposent une méthode pour trouver la présence d'une structure dans une classification. La méthode exploite une mesure de stabilité d'une classification basée sur la perturbation de l'ensemble des données. Ils utilisent la distribution des paires de similarités entre classifications des sous–ensembles de données comme mesure de stabilité d'une partition. Une grande valeur de paires de similarités indique une classification stable. Le nombre des classes est optimal lorsqu'un passage des solutions de classifications stables à des classifications non stables a eu lieu.

La netteté de ce passage peut donner une information sur la structure de données et fournir une information sur le manque de structure.

1.6 Conclusion

Nous venons de présenter dans ce chapitre les modèles probabilistes qui évaluent et étudient l'existence d'une partition, tels que les modèles de partitions fixes et les modèles de mélanges dont en particulier les modèles des classes latentes. Par la suite, c'est ce dernier qui sera utilisé pour générer nos partitions proches. Afin de réaliser les partitions, nous avons exploré quelques algorithmes de classifications tels que les k-means et les algorithmes ascendants. Les travaux traitant les problématiques de la validation et de la détermination du vrai nombre des classes d'une partition ont été évoqués. (Cette présentation n'est qu'un panorama servant à la compréhension de ce qui sera évoqué par la suite lors de la comparaison de deux partitions proches).

Chapitre 2

Interprétation des classes

2.1Introduction

De nombreuses méthodes pour l'interprétation des classes d'une partition d'un ensemble de données ont été proposées : nous en faisons ici une présentation synthétique.

Dans un premier temps, on présente les méthodes classiques utilisées en analyse de données qui se basent sur les caractéristiques des individus appartenant à une même classe à partir des valeurs ou modalités des variables tant actives que supplémentaires.

En deuxième lieu et dans le cadre de l'analyse des données symboliques, on s'intéresse aux travaux offrant une aide à l'interprétation des résultats, au moyen de règles logiques.

D'une part, l'utilisation des critères symboliques donne des résultats directement interprétables, et d'autre la méthode CABRO, proposé par H.T.Bao [BAO 88], produit un ensemble de règles aux classes qui fournissent les conditions d'appartenance d'un individu aux classes.

En troisième lieu, on présente la méthode de marquage sémantique proposée par M.Gettler- Summa [GET 93]. Son approche consiste à utiliser des indicateurs statistiques comme la valeur-test associée au test hypergéométrique d'égalité des proportions de M (marquage sémantique) dans une classe G et dans non-G, pour élaborer les descriptions caractérisant les classes. Ce type de description, appelé marquage sémantique, offre une règle logique fondée sur un indicateur statistique.

La dernière partie est consacrée à l'une des méthodes de classification divisives proposée par Marie Chavent [CHA 97]. Cette dernière est définie pour tous types de variables et une extension du critère d'inertie intra-classe. Elle est monothétique ce qui permet de

munir chaque classe de la hiérarchie d'une description simple facilement interprétable par l'utilisateur.

2.2 Méthodes classiques

Pour caractériser les classes d'une partition d'un ensemble des individus, on cherche les caractéristiques des individus appartenant à une même classe et cela en terme de variables. Plusieurs méthodes peuvent être utilisées :

2.2.1 Caractérisation unidimensionnelle des classes

Caractérisation par des variables illustratives

Les variables illustratives sont des variables qui ne contribuent pas à la construction des classes mais qu'on utilise *a posteriori* pour identifier et caractériser les regroupements établis à partir des variables actives. Pour déterminer les variables les plus caractéristiques de chaque classe, on compare la moyenne (ou la fréquence d'une modalité) d'une variable dans la classe à la moyenne (ou la fréquence) de cette variable dans l'échantillon total, en faisant l'hypothèse nulle H_0 que les individus qui constituent la classe sont tirés au hasard et sans remise dans l'échantillon global.

Caractérisation par une variable illustrative nominale

On s'intéresse à la variable aléatoire M qui représente le nombre d'individus de la classe C_h présentant la modalité j de la variable nominale. Sous H_0, cette variable suit une distribution hypergéométrique de moyenne :

$$E_h(M) = n_h . \frac{n_j}{n}$$

et de variance $$s^2_h(M) = n_h \frac{n - n_h}{n - 1} \frac{n_j}{n} \left(1 - \frac{n_j}{n} \right)$$

Où n est la taille de l'échantillon E global, n_h est l'effectif de la classe C_h et n_j est le nombre d'individus de E présentant la modalité j de la variable nominale.

Cette distribution peut être approximée par une distribution normale si les effectifs des classes sont assez élevés. Dans ce cas, $t_h(M) = \frac{M - E(M)}{s_h(M)}$ suit une loi normale centrée réduite.

H_0 : les nombres d'individus présentant la modalité j de la variable nominale dans la classe C_h et dans l'échantillon global sont égaux, aux fluctuations d'échantillon près. Ainsi le degré de significativité du test de H_0 est alors :

$$p_h(j) = P\left[|X| > t_h(n_{hj})\right]$$

Où X est la variable normale centrée réduite et n_{hj} est le nombre observé d'individus présentant la modalité j dans la classe C_h . Plus cette probabilité est faible, plus la modalité j est caractéristique de la classe C_h.

$t_h(n_{hj})$ est la valeur-test fournie par le logiciel SPAD. Elle représente l'écart entre la proportion dans la classe C_h et la proportion globale en nombre d'écart-type d'une loi normale.

Caractérisation par une variable illustrative continue

Sous la même hypothèse nulle H_0, les moyennes de la variable continue X dans la classe $C_h(\overline{X_h})$ et dans l'échantillon global $\overline{X}$ sont égales aux fluctuations aléatoires près. La valeur- test est alors :

$$t_h(X) = \frac{\overline{X_h} - \overline{X}}{s_h(X)} \qquad \text{avec} \quad s_h^2(X) = \frac{n - n_h}{n - 1}\frac{s^2(X)}{n_h}$$

$s^2(X)$ est la variance de la variable continue X calculée dans l'échantillon global.

Plus la valeur-test est grande, plus la variable est caractéristique de la classe.

Caractérisation par des variables actives

On peut calculer la valeur-test pour les variables actives mais dans ce cas, on considère ces valeurs-test comme des écarts entre les valeurs relatives à une classe et les valeurs de l'échantillon global. Elles permettent d'opérer un tri sur les variables continues et nominales et de désigner ainsi les variables les plus caractéristiques [LEB 97]. En effet

pour des variables actives, les propriétés distributionnelles ne sont plus exactes, car elles déterminent les classes.

2.2.2 Caractérisation multidimensionnelle des classes

Elle consiste à déterminer les variables qui caractérisent au mieux les classes obtenues, en les prenant en compte simultanément dans l'analyse. La variable classe, qui est ajoutée au tableau de données, joue le rôle de variable à expliquer dans une méthode d'analyse discriminante paramétrique ou non paramétrique. On peut effectuer l'une de ces trois méthodes :

- Une méthode de segmentation en prenant comme explicatives les variables initiales sans avoir à opérer un codage quelconque [BRI 84], [CEL 94].
- Une analyse discriminante (linéaire ou quadratique) à partir des composantes factorielles (variables explicatives) et remonter ensuite aux variables initiales.
- Une régression logistique en prenant comme variables explicatives les variables initiales (après codage si nécessaire) pour expliquer la variable classe.

2.2.3 Positionnement et dispersion des classes dans un plan factoriel

Le partitionnement en classes opère un découpage plus ou moins arbitraire d'un espace continu. L'analyse factorielle permet de visualiser sur un ou plusieurs plans factoriels significatifs, les positions relatives des classes. Dans ce but, on construit à l'issue d'une classification une variable nominale (de groupe) avec autant de modalités que de classes obtenues. Les modalités de cette variable de groupe ajoutée à la base de données agissent comme variables binaires illustratives.

Projetées en éléments supplémentaires sur un plan factoriel, elles fournissent les positions des points moyens des individus qui constituent une classe. On peut ainsi apprécier les distances entre classes. Par ailleurs, la position de chaque individu, repérée par le numéro de sa classe, permet de représenter la densité et la dispersion des classes dans le plan.

D'autre part, en ne gardant que les éléments (variables actives et illustratives) pertinents mis en évidence dans l'étape de caractérisation des classes, on simplifie la représentation graphique qui devient plus lisible.

Ces méthodes de caractérisation des classes peuvent être utilisées à l'aide des logiciels comme SPAD et SAS dont on pourra trouver des exemples illustratifs dans [NAK 00].

2.3 Analyse des Données Symboliques (ADS)

Les données réelles sont souvent plus complexes que la forme standard d'un tableau individus-variables où dans chaque case on ne trouve qu'une seule valeur. L'analyse de données symbolique développées par E. Diday [DID 91], est une nouvelle approche du traitement des connaissances. Son objectif est d'étendre la problématique, les méthodes de l'analyse de données à des données qui expriment un niveau de connaissances plus important que les simples observations.

Une description d'un individu peut être formée par plusieurs modalités ou par un intervalle de valeurs. Lors du traitement, ces valeurs sont gardées pour ne pas perdre l'information. L'analyse de données symbolique a également plusieurs autres objectifs :

- Définir un formalisme adapté pour le traitement des données complexes.
- Fournir des résultats qui sont munis d'une interprétation symbolique ou conceptuelle ou encore qui sont exprimés en termes de la description des individus.
- Définir de nouveaux critères symboliques.

Dans ce qui suit, on donne la définition d'un tableau de données symboliques, la description symbolique d'une variable, et les objets symboliques de types booléen et modale.

2.3.1 Tableau Individus-Variables en ADS

On désigne par individus, l'objet de l'analyse. Le terme individus peut se référer à une unité statistique (entité, ensembles d'entités) ou encore un concept.

Notons Ω l'ensemble des individus en entrée d'une analyse des données symboliques.

Dans le cadre de l'ADS, un tableau de données symboliques contient n individus w_i décrits sur p variables Y_j définie par :

Un espace des descriptions dont Δ_j est le domaine de description de la variable Y_j

Un ensemble O_j, appelé domaine des observables

Une structure algébrique S sur O_j

Une application[1] Y_j, définie sur Ω à valeurs dans Δ_j :

$$Y_j : \Omega \longrightarrow \Delta j$$

$$w \longrightarrow Y_j(w) = \delta_j$$

2.3.2 Type des variables

Plusieurs types des données symboliques ont été définis par Diday, on s'intéresse aux données [DID 96] et aux tableaux de données symboliques qui constituent la base de l'analyse de données symboliques. Ils sont définis de la manière suivante : les colonnes du tableau de données correspondent aux variables symboliques qui sont utilisées pour décrire l'ensemble des individus. Les lignes sont appelées les descriptions symboliques des individus car elles ne sont pas seulement des vecteurs de valeurs uniques. Une case du tableau symbolique peut être de différents types, en particulier :

- Valeur Quantitative: O_j est un ensemble continu et S admet une relation d'ordre $\leq$ ou c'est une structure de corps ordonné $\{\leq,+,*\}$. A titre d'exemple, si 'taille' est une variable et w est individu : taille(w)=1.9.
- Valeur Qualitative nominale: O_j est un ensemble dénombrable et S n'admet que la relation d'équivalence $\{=,\neq\}$. Exemple : ville(w)= paris.
- Ordinale: O_j est un ensemble dénombrable et S admet une relation d'ordre $\leq$. Exemple : niveau(w)=cadres.

Un ensemble de valeurs qualitatives (variable multivaluée). Par exemple, si la variable Y_j est les taxes payées par les trois plus grandes entreprises alors $Y_j(w)=\{900000, 750000, 500000\}$.

Un intervalle : on peut trouver une variable sous forme d'intervalle. Exemple taille(w)=[1,3], alors la taille de l'individu w varie selon l'intervalle [1,3].

[1] On donne le même nom à la variable et à l'application.

Ensemble de valeurs avec de poids associés : c'est dans le cas où la variable serait sous forme d'un histogramme, des fonctions d'appartenance ou bien des distributions de probabilités.

De plus, une variable peut avoir des relations logiques dont on peut trouver des connaissances supplémentaires. Ceci se traduit à deux niveaux :

Variable taxonomique : Une variable Y_j est taxinomique si O_j est un ensemble dénombrable muni d'une structure de graphe. Une arrête de ce graphe exprime une relation de généralisation entre deux modalités. On introduit un nœud racine, qui généralise toutes les valeurs observées. La spécialisation par l'utilisation de la taxinomie, est de correspondre à un ensemble de valeurs, les feuilles du plus petit sous arbre les contenant. La figure 2.1 montre un exemple sur une structuration hiearchique du domaine d'observation O_j={Employé, Cadres, Directeur, chef de département, Ingénieur} de la variable 'Niveau'.

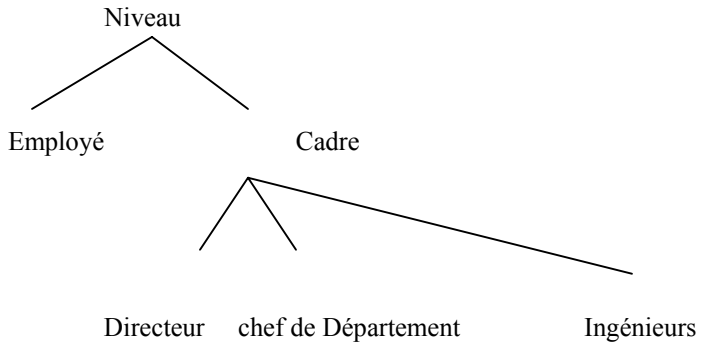


Fig. 2.1. *Exemple de taxinomie sur la variable 'Niveau'*

On parle parfois de variables structurées [MIC 83] et [ICH 94].

Relations entre variables : lorsqu'on exprime les liens connus entre les valeurs du domaine d'observation de certaines variables. On parle de relations entre variables. On distingue plusieurs types de relations ou de dépendances :

- **Dépendances logiques** [DEC 94] : Deux variables Y_1 et Y_2 sont liées par une dépendance logique si les valeurs du second variable dépendent logiquement ou fonctionnellement de la première variable. Cette dépendance est décrite par une règle illustrée dans l'exemple suivant : si Y_1=poids et Y_2=taille, il est possible qu'il y ait une règle de dépendance « r : si $Y_1 \leq 55$ alors $Y_2 \leq 180$ ».

- **Dépendances hiérarchiques** [LEB 91] (ou variable mère- fille) : Une variable Y_1 dépend hiérarchiquement d'une autre variable Y_2, si Y_1 est conditionné par les valeurs observées sur Y_2, dans ce cas Y_2 est la variable mère et Y_1 est la variable fille. Les deux variables sont liées par une règle de non-applicabilité. A titre d'exemple, la règle « Voiture= non donc consommation d'essence=NA », indique que la consommation d'essence est non applicable pour les individus qui n'ont pas de voitures. La valeur NA distingue la valeur qui n'est pas renseignée de la valeur qui ne peut pas être renseignée.

- **Dépendances stochastiques** [BOC 99] : les p variables aléatoires $Y_1,\ldots,Y_p$ sont indépendant stochastiquement sous la distribution de probabilité P si :

$$P(Y_1 \in B_1,....,Y_p \in B_p) = \prod_{j=1}^{p} P(Y_j \in B_j)$$

pour tout sous ensemble mesurable $B_j \subset \mathcal{Y}_j,\ \forall j = 1,...,p$

Notons que cette formule est équivalente à celle écrite par un produit de densité et de fonctions de probabilités.

2.3.3 Types de données

D'une manière générale, les données symboliques d'une variable Y_j distinguent le domaine d'observation O_j du domaine d'arrivée Δ_j. Ces données peuvent être classifiées selon les trois cas suivants :

$Y_j(w_i)$ est dite **donnée univaluée** si et seulement si : $\Delta_j = O_j$. C'est le cas des variables des valeurs uniques classiques.

$Y_j(w_i)$ est dite **donnée multivaluée** si et seulement si : $\Delta_j = P(O_j)$, les descriptions symboliques peuvent être des ensembles de valeurs ou des intervalles de valeurs.

Soit $O_j=\{1,2,\ldots,m\}$ et $card(O_j)=m$. $Y_j(w_i)$ est dite **donnée modale** si et seulement si :

$\Delta_j = [0,1]^m$. Δ_j peut être réduit à l'ensemble des distributions de probabilités, des distributions de fréquences ou de poids de O_j.

2.3.4 Les Opérateurs sur des descriptions complexes

Opérateur d'union et d'intersection

On ne s'intéresse qu'à la définition des opérateurs d'union et d'intersection dans le cas des données univaluées et multivaluées [HIL 98]. On utilisera ces opérateurs pour définir les fonctions de comparaison de deux individus sur une variable. En effet, deux descriptions sont similaires d'autant plus que leur intersection est grande et leur union est petite**.**

- **Définition de l'union**

On définit l'union sur P(Δ), comme une application :

$$\begin{aligned} &U: P(\Delta) \to P(O_1) \times P(O_2) \times ... \times P(O_p) \\ &\quad E \quad \to d \end{aligned}$$

où $d = (d_1, d_2, ..., d_p)$ est tel que $d_j = \bigcup(\{\delta_j / \delta \in E\}), j = 1,...,p$

Si Y_j est qualitative : $\bigcup(\{\delta_j / \delta \in E\})$ est l'union ensembliste des δ_j, c'est à dire dans le cas où les données seraient univaluées $d_j = \{v \in O_j / \exists \delta \in E, \delta_j = v\}$ et dans le cas où les données seraient multivaluées $d_j = \bigcup_{\delta \in E} \delta_j$

si Y_j quantitative : soit $\delta_j^1 = [\underline{\delta_j^1}, \overline{\delta_j^1}]$ et $\delta_j^2 = [\underline{\delta_j^2}, \overline{\delta_j^2}]$ on obtient :

$\delta_j^1 \cup \delta_j^2 = \{\delta_j^1, \delta_j^2\}$, si δ_j^1 et δ_j^2 sont des intervalles disjoints

$\delta_j^1 \cup \delta_j^2 = [\min(\underline{\delta_j^1}, \underline{\delta_j^2}), \max(\overline{\delta_j^1}, \overline{\delta_j^2})]$, sinon

- **Définition de l'intersection**

On définit cet opérateur d'intersection dans le cas où les données seraient multivaluées car il a peu de sens lorsque les données sont univaluées.

On définit l'intersection comme une application :

$$\begin{aligned} &\cap: P(\Delta) \to P(O_1) \times P(O_2) \times ... \times P(O_p) \\ &\quad E \quad \to d \end{aligned}$$

où $d = (d_1, d_2, ..., d_p)$ est tel que $d_j = \bigcap(\{\delta_j / \delta \in E\}), j = 1,...,p$

de même l'opérateur d'intersection est ensembliste :

si Y_j qualitative : $\bigcap(\{\delta_j / \delta \in E\})$ est l'intersection des δ_j,

si Y_j quantitative : soit $\delta_j^1 = [\underline{\delta_j^1}, \overline{\delta_j^1}]$ et $\delta_j^2 = [\underline{\delta_j^2}, \overline{\delta_j^2}]$ on obtient :

$\delta_j^1 \cap \delta_j^2 = \phi$, si δ_j^1 et δ_j^2 sont des intervalles disjoints

$\delta_j^1 \cap \delta_j^2 = [\max(\underline{\delta_j^1}, \underline{\delta_j^2}), \min(\overline{\delta_j^1}, \overline{\delta_j^2})]$, sinon.

- Définition de l'union symbolique ou de la jonction

L'union de deux intervalles n'est pas toujours un intervalle, par exemple : [1,3] U [5,8]. Le terme de jonction a été proposé par M. Ichino [ICH 94] dans le but d'avoir toujours un intervalle lorsqu'on cherche l'union de deux intervalles.

Il définit l'union jointe dans le cas où le domaine d'observation serait muni de données multivaluées et d'une structure arborescente.

Pour de données multivaluées :

La jonction $\oplus$ entre deux descriptions de Δ est définie par :

$$\oplus : P(\Delta) \to P(O_1) \times P(O_2) \times ... \times P(O_p)$$
$$E \to d$$

où $d = (d_1, d_2, ..., d_p)$ est tel que $d_j = \oplus(\{\delta_j / \delta \in E\}), j = 1,...,p$

L'opérateur $\oplus$ sur $P(O_j)$ est défini comme :

$$\oplus(\{\delta_j / \delta \in E\}) = \begin{cases} \bigcup_{\delta \in E} \{\delta_j\} & \text{si } Y_j \text{ non ordonnée} \\ [\min(\{\underline{\delta_j} / \delta \in E\}), \max"(\{\overline{\delta_j} / \delta \in E\})] & \text{sinon} \end{cases}$$

L'opérateur de jonction est identique à l'union dans le cas de données portant sur des variables qualitatives. Dans le cas de données intervalles, il permet d'obtenir une description plus générale.

Pour une structure d'arbre :

Soient δ_A et δ_B sont deux éléments de $P(O_j)$ et le domaine d'observation O est nominal et structuré dans une hiérarchie. Si $N(\delta_A)$ est la classe la plus fine dans la hiérarchie, qui contienne tous les éléments de δ_A, alors $\delta_A \oplus \delta_B$ est défini par :

$$\delta_A \oplus \delta_B = \begin{cases} \delta_A \cup \delta_B & \text{si } N(\delta_A) = N(\delta_B) \\ N(\delta_A \cup \delta_B) & \text{sinon} \end{cases}$$

Par exemple, la variable 'branche' dont le domaine d'observation {droit, langue, physique, chimie, informatique, statistique} est une variable taxonomiques (Fig. 2.2).

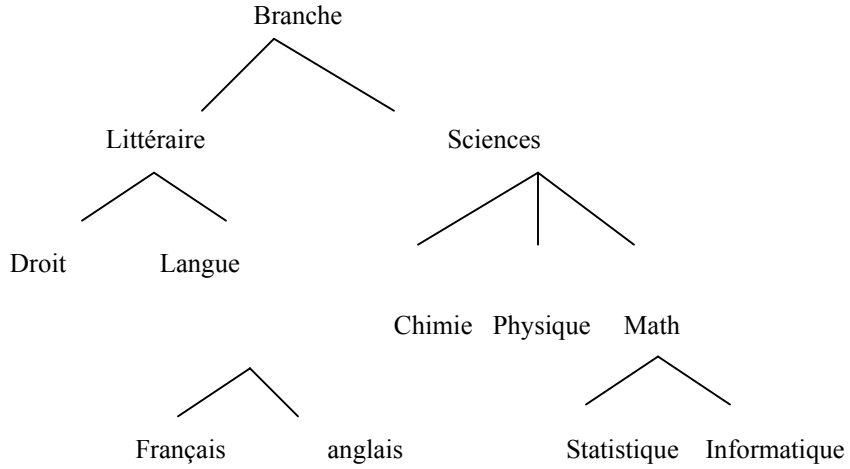


Fig. 2.2. *Exemple de taxinomie sur la variable 'branche'*

Si δ_A ={Math, Physique} et δ_B ={Chimie, Physique} alors $N(\delta_A)$=$N(\delta_B)$ = 'Sciences'

Si δ_A ={Droit} et δ_B ={Français} dans ce cas $N(\delta_A)$= Littéraire et $N(\delta_B)$= Langue, donc $N(\delta_A \cup \delta_B)$=Littéraire et par suite $\delta_A \oplus \delta_B$ ={Droit, Français, Anglais}.

2.3.5 Présentation des Objets Symboliques

L'ADS offre une représentation explicite des connaissances. On distingue deux niveaux :

- Un niveau concernant les objets traités en entrée d'une analyse (les individus ayant une description symbolique), d'où la description individuelle. On cherche une description d d'une classe C d'individus.
- Un niveau concernant les objets obtenus en sortie d'une analyse (les classes ayant une description symbolique), d'où la description intentionnelle. Si on commence

par une description d, c'est important de savoir tous les individus qui sont constitués par d.

Plusieurs auteurs ont donné une définition à l'objet symbolique comme étant un vecteur de descriptions symboliques, en lui associant une représentation logique, dont l'interprétation permet d'obtenir son 'extension'[PER 96], [BRI 91], [DEC 92].

Un objet symbolique s'exprime sous la forme d'une conjonction de plusieurs propriétés des variables de l'analyse.

On définit deux types d'objet symbolique :

Les **assertions** : C'est le type le plus utilisé en ADS. Les assertions sont de descriptions adaptées au tableau de données. A titre d'exemple, une classe C de descriptions connues peut être représentée par l'assertion suivante :

a(C)= [age(C)=[50,70]] ^[nombres d'enfants(C)={0,1}]

Les **hordes** : ils sont des expressions logiques qui prennent en compte la structuration des parties décrivant un même objet. Par exemple, si P est une partition de deux classes C_1 et C_2, on peut décrire P par la horde suivante :

H(P)= [age(C_1)=[50,70]] ^[nombres d'enfants(C_1)={0,1}] ^[age(C_2)=[40,50]]^[nombres d'enfants(C)={2,3}]

Alors par l'utilisation des objets symboliques, les sorties d'une analyse s'interprète facilement. On obtient une description conceptuelle. Un concept est défini en 'intension' ou en 'extension'. L'intension d'un concept définit un ensemble de propriétés qui décrit le concept. Un individu appartient à un concept s'il satisfait aux conditions formées à partir de ces propriétés. Cet individu est alors une instance du concept. La liste de ses instances s'appelle extension du concept.

Formalisme des objets symboliques

Le formalisme des objets symboliques, proche du formalisme logique adopté dans le cadre de l'apprentissage symbolique automatique, est fondé sur deux idées duales :

Fournir à l'utilisateur une représentation explicite résumant les observations d'un ensemble d'individus. Par exemple, dans le cas d'une méthode de partition, on associe

aux classes de la partition, un ensemble de descriptions généralisant les éléments de chacune des classes. Considérons $C \in P(\Omega)$, le vecteur $d \in D$ tel que d=g(C) ;

Associer à la description d, une application notée a, permettant de calculer le degré d'adéquation d'un individu ω à d. On identifie l'application a à une relation de subsomption [DEG 90], [NAP 92] d'un individu sous un concept.

Pour définir un objet symbolique *s* on définit une description *d* (intensionnelle en générale), d'une relation *R* entre les descriptions, et d'une fonction a :

$$a : \Omega \rightarrow \Delta$$
$$\omega \rightarrow a(\omega) = [\, Y(\omega) R d]$$

Cette fonction évalue le degré d'appartenance d'un individu ω à l'extension de *s.*

Un objet symbolique *s* est un triplet *s=(a, R, d)* où d est le vecteur de description, la relation entre les descriptions $R \in \{\subset, \supset, \subseteq, \supseteq\}$.

Parmi les objets symboliques on s'intéresse aux assertions. Une assertion s est un couple (d, a), où d est la généralisation d'un élément C. Une assertion est exprimée sous la forme d'une conjonction de termes :

$$a = \wedge_j [Y_j \in d_j]$$

On distingue deux types d'objets symboliques : objets symboliques booléens et objets symboliques modales.

Objets symboliques booléens : on commence par un tableau de p variables de différents types de terme général $Y_j(\omega)$ pour un individu ω, d'une description d muni d'une relation binaire R. Les valeurs de a peuvent être vraie ou faux.

Une assertion booléenne s, est un couple s=(d,a), où a est une application définie à valeur dans $\{0,1\}$.

un exemple d'assertion booléenne de deux variables Y_1=age, et Y_2=niveau de travail, est donnée par :

$$a(\omega) = [age(\omega) \subseteq \{20,30,50\}] \wedge [niveau(\omega) \subseteq \{employé, cadre\}]$$

Dans ce cas, on a une syntaxe d'objets symboliques illustrée de la façon suivante :

Si dans le tableau symbolique de base, un individu ω est décrit par :

$$\text{age}(\omega)=\{20,30\} \text{ et niveau}(\omega)=\{\text{employé}\}$$

Donc on a $d_1=\{20,30,50\}$ et $d_2=\{$ employé, cadre$\}$, $d=\{d_1,d_2\}$ et $R_j=\subseteq$ pour j=1,2.

$a(\omega) = [Y(\omega)Rd] = [\{20,30\} \subseteq \{20,30,50\}] \wedge [\{ \text{employé}\} \subseteq \{ \text{employé, cadre}\}] =$ vraie $\wedge$ vraie = vraie.

On pourra utiliser l'opérateur logique de disjonction $\vee$ pour trouver des objets symboliques booléens. D'une façon similaire à l'assertion : $a(\omega) = \vee [Y(\omega)Rd]$.

L'extension d'une assertion booléenne s=(d,a) est l'ensemble des individus de Ω ayant $a(\omega)$= vraie. Elle est identique à l'extension de a, on a :

$$\text{ext}_{\Omega}(s)=\text{ext}(a)=\{\omega \in \Omega / a(\omega) = \text{vraie}\}.$$

Objets symboliques modales : De la même manière, on appelle objets symboliques modales lorsque la fonction a prend ses valeurs dans l'intervalle [0,1]. Dans ce cas, le choix de la relation R est lié à la sémantique associée à la description de l'individu. Les valeurs prises par la variable Y_j sont distribuées selon une loi de probabilité.

Par exemple, si on a la variable 'age' est normalement distribuée, et on a :

$$a(\omega) = [\text{age}(\omega) \in [20,50]]$$

On peut définir $a_s(\omega)$ comme étant la probabilité pour que son age appartient à l'intervalle [20,50].

L'extension d'un objet modal peut être définie de deux manières différentes [DID 92] : On considère tout d'abord que tout individu $\omega \in \Omega$ peut appartenir « plus ou moins » à l'extension de s, en fonction de son degré d'appartenance $a(\omega)$:

$$\text{ext}_{\Omega}(s)=\{(\omega, a(\omega)) / \omega \in \Omega\}$$

Dans une deuxième approche, on peut considérer que l'appartenance de l'individu ω à l'extension de s est admise si la quantité $a(\omega)$ est au moins égale à un seuil α fixe :

$$\text{ext}_{\Omega}(s)=\{\omega \in \Omega / a(\omega) \geq \alpha\}$$

2.3.6 Méthode « CABRO » et les Critères Symboliques

Méthode CABRO

La méthode CABRO, proposée par Ho Tu Bao [BAO 88], construit un ensemble de règles associées aux classes d'une partition. Cette méthode a été améliorée [BAO 91] pour déterminer la caractérisation des classes selon des règles plus générales et basées non seulement sur la classe exemple mais sur la classe complémentaire d'une partition. Cette méthode construit, à partir d'un ensemble de données, une règle de base pour la caractérisation des classes. Ces règles peuvent être utilisées pour reconnaître les individus dans un domaine particulier.

Exemple:

Forme(w)= bombée ^ couleur(w) = rouge ^ taille (w)= grande $\Rightarrow w \in C$

Cette règle offre une caractérisation de la classe C. Elle indique les modalités les plus significatives et les plus spécifiques observées sur les éléments de C et qui la distingue des autres classes. Chaque classe de la partition peut être caractérisée par une ou plusieurs règles. L'ensemble des règles, associées à une même classe, fournit les conditions nécessaires et suffisantes pour l'appartenance d'un individu à celle-ci. Mais, il faut noter que la qualité des résultats dépend de l'ensemble de données initiales.

Une adaptation de la méthode CABRO, a été proposée par M. Gettler- Summa [GET 94], pour trouver des caractérisations des classes d'une partition d'un ensemble d'objets symboliques modales. Les assertions des classes ne sont pas des critères de généralisations optimaux, mais elles sont basées sur le choix de variables demandés.

Utilisation de Critères symboliques

L'avantage de l'utilisation de critères fondés sur des propriétés logiques est d'obtenir des résultats directement interprétables. Le premier travail introduisant un critère symbolique est celui de la classification conceptuelle proposée par E. Diday, [MIC 81], concernant une adaptation des nuées dynamiques. Leur technique classificatoire repose sur la recherche combinée de classes disjointes d'individus et de leur caractérisation. On associe à chaque classe une description sous la forme d'une conjonction de termes ou chaque terme modélise la variation des observations au sein de la classe pour une même

variable. Le critère symbolique proposé est une combinaison de critères comme la simplicité qui réfère au nombre de termes de chaque description ou encore le taux de mauvais classement déterminé à partir des descriptions caractérisantes.

2.4 Marquage Sémantique

Le marquage sémantique des classes et des axes factoriels [MOR 95] est une nouvelle aide à l'interprétation des classes obtenues par la classification automatique. Les résultats s'expriment en termes d'assertion construite à partir des libellés des paramètres décrivant les individus.

On considère un groupe fixe d'individus, sous-ensemble de l'échantillon global. On cherche à répondre à la question suivante : qu'est ce qui distingue les individus de ce groupe de ceux qui ne sont pas dans le groupe ?

Pour décider si un attribut (ou une modalité d'une variable) est une caractéristique du groupe, on teste l'égalité des proportions de l'attribut dans le groupe et dans son complémentaire. L'attribut est caractéristique si l'hypothèse d'égalité peut être rejetée. Le test se ramène au calcul d'une probabilité hypergéométrique. La probabilité critique du test [MOR 84] est utilisée pour classer les attributs par ordre d'intérêt décroissant dans la caractérisation du groupe.

Toutes les combinaisons de tous les attributs, utilisant les opérateurs ET et OU, sont candidates à la caractérisation du groupe. La procédure de marquage sémantique a pour objet de rechercher, parmi toutes les propositions construites par réunions et intersections d'un nombre quelconque d'attributs, celles qui sont le plus caractéristique du groupe.

2.4.1 Présentation de l'algorithme

On considère une population P et un groupe G d'individus : G est un sous-ensemble fixe de P. Un marquage sémantique est l'un des sous-ensembles M de P. Le marquage sera caractéristique du groupe G s'il recouvre bien G et s'il déborde peu de G.

Le critère utilisé pour quantifier l'intérêt du marquage M dans la caractérisation de G est la valeur-test associée au test d'égalité des proportions de M dans G et dans non G. L'objectif de l'algorithme est de trouver le meilleur marquage, c'est à dire le compromis

optimal du grand recouvrement de G pour le plus petit débordement dans non G. On décrira cet algorithme par un exemple extrait des données SPAD ;

Exemple

L'exemple est pris de la base de données SPAD qui s'appelle 'enquête'. On dispose de 315 individus, et de 52 variables dont 40 d'entre elles sont nominales. On s'intéresse au paramètre 'sexe' formé de 2 modalités : 177 femmes et 138 hommes.

Considérons le groupe des 177 femmes. Pour caractériser le groupe, la procédure détermine 4 marquages, chacune est déterminée par des conjonctions différentes :

```
MARQUAGE NUMERO     1             POIDS =       52.0    POURCENTAGE =  16.51%
-------------------------------------------------------------------------------------------------------
                  RECOUVREMENT       RECOUVREMENT       RECOUVREMENT       DEBORDEMENT
     V.TEST     PAR LE MARQUAGE         AJOUT              CUMUL           DU MARQUAGE
                 POIDS       %      POIDS       %      POIDS       %      POIDS       %
       7.87       52.0    29.4       52.0    29.4       52.0    29.4        0.0     0.0
                MODALITE             LIBELLE DE LA VARIABLE
       7.87     ménagère s.prof.      Situation actuelle de la personne interrogée
-------------------------------------------------------------------------------------------------------
MARQUAGE NUMERO     2             POIDS =       26.0    POURCENTAGE =   8.25%
-------------------------------------------------------------------------------------------------------
                  RECOUVREMENT       RECOUVREMENT       RECOUVREMENT       DEBORDEMENT
     V.TEST     PAR LE MARQUAGE         AJOUT              CUMUL           DU MARQUAGE
                 POIDS       %      POIDS       %      POIDS       %      POIDS       %
       4.51       25.0    14.1       17.0     9.6       69.0    39.0        1.0     3.8
                MODALITE             LIBELLE DE LA VARIABLE
       3.51     oui                   Avez-vous souffert récemment de nervosité
 et    1.88     dissout si pb. grave  Opinion sur le mariage
 et    1.70     beaucoup              Etes-vous inquiet du risque d'une maladie grave
 et    1.70     non                   Disposez-vous d'une résidence secondaire
-------------------------------------------------------------------------------------------------------
MARQUAGE NUMERO     3             POIDS =       19.0    POURCENTAGE =   6.03%
-------------------------------------------------------------------------------------------------------
                  RECOUVREMENT       RECOUVREMENT       RECOUVREMENT       DEBORDEMENT
     V.TEST     PAR LE MARQUAGE         AJOUT              CUMUL           DU MARQUAGE
                 POIDS       %      POIDS       %      POIDS       %      POIDS       %
       4.24       19.0    10.7       14.0     7.9       83.0    46.9        0.0     0.0
                MODALITE             LIBELLE DE LA VARIABLE
       4.24     veuf(ve)              Statut matrimonial
-------------------------------------------------------------------------------------------------------
MARQUAGE NUMERO     4             POIDS =       31.0    POURCENTAGE =   9.84%
-------------------------------------------------------------------------------------------------------
                  RECOUVREMENT       RECOUVREMENT       RECOUVREMENT       DEBORDEMENT
     V.TEST     PAR LE MARQUAGE         AJOUT              CUMUL           DU MARQUAGE
                 POIDS       %      POIDS       %      POIDS       %      POIDS       %
       4.11       28.0    15.8       15.0     8.5       98.0    55.4        3.0     9.7
                MODALITE             LIBELLE DE LA VARIABLE
       1.78     non                   La famille est le seul endroit où l'on se sent bien
 et    1.74     tous les jours        Regardez-vous la télévision ?
```

Tab.2.1. *Résultat par la méthode du marquage sémantique en SPAD*

Le premier marquage est défini par la modalité «`ménagère s. prof`». Il y a 52 femmes qui sont ménagères et sans profession. Ces femmes recouvrent 29,4 % du groupe. L'abondance de ce marquage dans le groupe à caractériser est quantifiée par la valeur-test 7.87. La probabilité critique du test d'égalité des proportions des individus marqués

dans le groupe et hors du groupe correspond à une distance de 7.87 écarts- types sur l'échelle normale.

Le deuxième marquage est défini par la conjonction « souffert récemment de nervosité », « dissout si problème grave », « inquiet du risque d'une maladie grave », et « ne dispose pas d'une résidence secondaire ». Il y a 25 femmes qui recouvrent ce marquage dont 17 ne sont pas observées par le premier marquage et une qui est hors du groupe. Les deux marquages recouvrent 39 % du groupe. Les autres marquages se lisent de la même façon.

On remarque l'aspect sémantique de la caractérisation : le groupe est caractérisé par les différentes assertions qui sont définies par la conjonction des modalités présentes.

2.5 Méthodes Divisives de classification

Les méthodes divisives de classification sont des méthodes hiérarchiques descendantes qui obtiennent par division successive de l'ensemble des individus une suite de partitions emboîtées. Ces méthodes fournissent une interprétation simple des classes de la hiérarchie.

Les méthodes divisives de type monothétiques ont la particularité de construire des classes dont les individus vérifient un ensemble de propriétés nécessaires et suffisantes d'appartenance à la classe.

M.Chavent [CHA 99] a développé une méthode divisive pour les données classiques et symboliques. On présente la méthode dans le cas de données classiques, multivaluées, sous forme d'intervalle ou modales. On cherche à chaque étape à diviser la classe qui fournit une nouvelle partition optimisant sous contrainte un critère mathématique choisi. Les contraintes sont induites par l'aspect monothétique de la méthode et le critère à optimiser peut être une extension du critère d'inertie aux cas de données de descriptions symboliques.

La division de la classe en deux autres consiste à minimiser la variance intra-classe en respectant les partitions induites par les ensembles des questions binaires relatives à toute variable.

2.5.1 Présentation de la méthode

Considérons un ensemble de n objets caractérisés par p variables symboliques.

$$Y_j : \Omega \longrightarrow \Delta_j$$
$$k \longrightarrow Y_j(k)$$

Pour les valeurs réelles, on considère deux types :

- Cas classique de variables quantitatives : $Y_j(k) \in IR$, avec $\Delta_j = IR$
- Cas de variables sous forme d'intervalle : $Y_j(k) = [\alpha, \beta] \subset IR$, donc Δ_j sera l'ensemble des intervalles borné fermé dans IR

Pour les valeurs qualitatives ordonnées $Y_j = \{a,b,c,\ldots,h\}$, on considère trois types de variables :

- Cas de variables ordinales classiques : $Y_j(k) \in Y_j$ et $\Delta_j = Y_j$
- Cas de variables multivaluées : $Y_j(k) \subset Y_j$ et et $\Delta_j = P(Y_j)$
- Cas de variables modales : $Y_j(k) = \pi_k$ une probabilité sur Y_j et Δ_j est l'ensemble de toutes les probabilités sur Y_j

Pour déterminer la matrice des distances, on considère deux distances différentes:

Distance d'une matrice de données symboliques quantitatives

On fait la combinaison de p indices de dissimilarités définis sur Δ_j, en utilisant la distance de Hausdorff :

Si ξ_{kj} et ξ_{lj} $[\alpha_k^j, \beta_k^j]$ et $[\alpha_l^j, \beta_l^j]$ sont deux intervalles, la distance est :

$$d_j(\xi_{kj}, \xi_{lj}) = \max\{|\alpha_k^j - \alpha_l^j|, |\beta_k^j - \beta_l^j|\}$$

La fonction de distance d qui combine ces indices est alors :

$$d : \Omega \times \Omega \rightarrow IR^{+}$$

$$(k,l) \rightarrow d(k,l) = \left(\sum_{j=1}^{p} d_j(\xi_{kj}, \xi_{lj})^2 \right)^{1/2}$$

Distance d'une matrice des données symboliques qualitatives

Dans le cas des données qualitatives, pour comparer deux objets *k* et *l* de Ω, on utilise la distance Φ^2 :

$$d^2(k,l) = \sum_{j=1}^{p} \frac{p_{..}}{p_{.j}} \left(\frac{p_{kj}}{p_{k.}} - \frac{p_{lj}}{p_{l.}} \right)^2$$

avec $p_{..} = \sum_{k=1}^{n} \sum_{j=1}^{t} p_{kj} = 1$; $p_{k.} = \sum_{k=1}^{n} p_{kj}$; $p_{.j} = \sum_{j=1}^{p} p_{kj}$;

Le critère utilisé pour évaluer la qualité d'une partition est une extension du critère de somme de carrée intra-classe pour une classe C_i

$$I(C_i) = \sum_{k \in C_i} \left\| x_k - \bar{x}_{C_i} \right\|^2$$

2.5.2 Bipartitionnement d'une classe

Le but est de trouver une bipartition $C_i=(C_i^1, C_i^2)$ de la plus petite variance intra-classe. Dans la méthode de classification divisive, la classe C est partagé selon la question suivante : Est-ce que $Y_j \leq c$**?**

En effet, dans le cas des données symboliques, un objet k de C répondant oui ou non à la question divise la classe en deux :$C_1=\{k \in C / q_c(k) = vrai\}$ et $C_2=\{k \in C / q_c(k) = faux\}$

avec $q_c : \Omega \rightarrow \{vrai, faux\}$ et c est valeur de coupure.

* Pour une variable sous forme d'intervalle :

$$q_c(k) = \begin{cases} \text{vrai} & \text{si } m_k \leq c \\ \text{faux} & \text{sinon} \end{cases}$$

où $c = \frac{\alpha + \beta}{2}$ pour $Y_j=[\alpha, \beta]$

* Pour une variable modale :

$$q_c(k) = \begin{cases} vrai & \text{si } \sum_{x \le c} \pi_k(x) \ge 1/2 \\ faux & \text{sinon} \end{cases}$$

Exemple

On applique la méthode divisive de classification en utilisant le logiciel SODAS sur un exemple à 100 individus et 21 variables. On sélectionne deux variables :

- Le sexe qui est divisé en deux modalités : homme et femme,
- Le résultat de la deuxième année des personnes demandées qui est aussi divisé en deux modalités : recalé et admis.

On présente l'arbre hiérarchique (Fig. 2.3) de classification des variables sélectionnées. Le nombre noté à chaque nœud indique l'ordre de la division des classes. Ng et Nd représentent respectivement l'accord et le désaccord.

```
          +----  Classe 1 (Ng=19)
          !
     !----3- [résultat de la 2eme année = recalé]
     !    !
     !    +----  Classe 4 (Nd=19)
     !
!----1- [sexe = homme]
     !
     !    +----  Classe 2 (Ng=31)
     !    !
     !----2- [résultat de la 2eme année = recalé]
          !
          +----  Classe 3 (Nd=31)
```

Fig.2.3. *L'arbre hiérarchique selon les deux variables 'sexe' et résultat de la 2^{ieme} année.*

L'arbre hiérarchique nous indique les 4 classes présentées selon les modalités de deux variables, la classe 1 est formée de 19 étudiants recalés en 2^{ième} année, la classe 2 est formée par 31 étudiantes recalées en 2^{ième} année, la classe 3 est formée par les étudiantes admises en 2^{ième} année, et la classe 4 est formée par 19 étudiants admis en 2^{ième} année.

2.6 Conclusion

On a pu définir les principales méthodes de classification qui sont capables de traiter des données de tous les types de variables et de fournir une interprétation des classes obtenues. On a présenté les méthodes classiques utilisées en analyse de données qui se basent sur les caractéristiques des individus appartenant à une même classe. Les travaux qui offrent une aide à l'interprétation des résultats, au moyen de règles logiques, ont été présentés: la méthode CABRO, la méthode de marquage sémantique, et la méthode de classification divisives [CHA 97]. L'application sur des données réelles des deux dernières méthodes d'interprétation des classes a été effectuée.

Chapitre 3

Indices de comparaison de deux partitions sur les mêmes individus

3.1 Introduction

Quand on dispose de deux partitions effectuées sur les mêmes individus, par exemple avec deux jeux de variables ou bien avec deux algorithmes, il faut savoir si ces deux partitions sont en accord ou bien si elles diffèrent significativement, en un sens à préciser. Une manière d'aborder ce problème consiste à calculer un indice de concordance entre partitions et à définir une valeur critique à partir de laquelle on conclura que les deux partitions sont ou non concordantes.

Ce chapitre est consacré à la présentation et à la définition des différents indices qui nous paraissent important dans notre propos. La plupart de ces indices sont présentés en formulations contingentielles et relationnelles en utilisant les formules de passages proposées par Kendall [KEN 61] et Marcotorchino [MAR 84].

A l'indice bien connu de Rand et celui corrigé par Hubert [HUB 85], on propose une version asymétrique de Rand [CHAV 01] utilisée pour la comparaison de partitions emboîtées, avec des nombres différents de classes. On ajoute deux autres indices inspirés de test de Mac Nemar et de l'indice de Jaccard. On présente l'indice de corrélation vectorielle introduit par P. Robert et Y. Escoufier [ROB 76] qui se révèle identique au coefficient de S. Janson et J. Vegelius [JAN 82], le coefficient kappa de Cohen [COH 60], l'indice de redondance proposé Stewart et Love [STE 68], ainsi que l'indice de Popping [POP 83].

L'utilisation des mesures habituelles d'associations comme le khi-deux ne permet pas de répondre de manière adéquate à la comparaison de partitions : mesure de l'écart à

l'indépendance, le khi-deux n'est pas adapté au problème qui consiste à tester l'écart à une structure diagonale ; L'hypothèse d'indépendance est inintéressante dans notre étude. C'est pourquoi l'indice de khi-deux ne sera pas considéré dans ce chapitre.

Ces travaux ont fait l'objet de communications dans des congrès et de publications [SAP 01, 02] et [YOU 03, 04, 04'].

3.2 Notations et définitions élémentaires

Dans ce paragraphe, nous introduisons les notations de bases, ainsi que les définitions élémentaires en classification qui seront utilisées.

P_1 et P_2 sont deux partitions des mêmes individus (ou deux variables qualitatives). N désigne le tableau de contingence associé, K_1, K_2 les tableaux disjonctifs associés à P_1 et P_2; On a : $\mathbf{N} = \mathbf{K_1'K_2}$.

Chaque partition P_k est représentée par un tableau relationnel C^k dans l'espace des individus, de dimension nxn, dont le terme général $c^k_{ii'}$ est défini par :

$$c^k_{ii'} = \begin{cases} 1 & \text{si i et i' sont deux individus dans la même classe de la partition } P_k \\ 0 & \text{sinon} \end{cases}$$

L'écriture matricielle du tableau de comparaison par paire est $C_1=K_1K_1'$. Au tableau C est associé son tableau complémentaire, notée $\overline{C}$ dont le terme général est défini par :

$$\overline{c}^k_{ii'} = 1 - c^k_{ii'} = \begin{cases} 0 & \text{si i et i' sont deux individus dans la même classe de la partition } P_k \\ 1 & \text{sinon} \end{cases}$$

Pour n nombre d'individus, on a p classes de la partition P_1 et q classes de la partition P_2. Lorsque l'on croise deux partitions, on va s'intéresser aux paires d'individus qui restent ou ne restent pas dans les mêmes classes. On a $\binom{n}{2}$ paires d'individus représentées par les quatre types dans le tableau suivant :

$P_1 \setminus P_2$	**Même classe**	**Classes différentes**
Même classe	a Accord positif	d Désaccord
Classes différentes	c Désaccord	b Accord Négatif

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

Tab 3.1 *Tableau croisant les deux partitions P_1 et P_2*

On notera également A=a+b (nombre total d'accords) et D=c+d (nombre total de désaccords). On peut aussi, au lieu de considérer les $\binom{n}{2}$ paires (i, i') considérer les n^2 paires (i, i') (où (i, i') est distingué de (i', i) et où l'on comptabilise les n paires (i, i)). Si a', b', c', d' désignent les équivalents de : a, b, c, d, on a alors :

$$a'=2a+n \; ; \; b'=2b \; ; \; c'=2c \; ; \; d'=2d$$

3.3 Formules de linéarisation

Le tableau de contingence croisant P_1 et P_2 est de dimension pxq. Il est caractérisé par son terme général : n_{uv}= l'effectif de la case (u, v), il est lié aux termes généraux des tableaux disjonctifs associées à P_1 et P_2 par la formule suivante :

$$n_{uv} = \sum_{i=1}^{n} k_{iu} k_{iv}$$

Les tableaux relationnels C^1 et C^2 fournissent la même information que le tableau de contingence croisant les partitions P_1 et P_2.

Pour notre étude, nous utilisons les formules de passages contingences-paires qui ont été proposées et démontrées par Kendall [KEN 61] et Marcotorchino [MAR 84] :

$$\sum_i \sum_{i'} c^1_{ii'} = \sum_u n^2_{u.}$$

$$\sum_i \sum_{i'} c^2_{ii'} = \sum_v n^2_{.v}$$

$$\sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} = \sum_u \sum_v n^2_{uv}$$

Ces formules nous ont permis d'établir les relations d'équivalences suivantes :

$$a' = 2a + n = \sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} = \sum_u \sum_v n^2_{uv}$$

$$b' = 2b = \sum_i \sum_{i'} \bar{c}^1_{ii'} \bar{c}^2_{ii'} = \sum_i \sum_{i'} (1 - c^1_{ii'})(1 - c^2_{ii'}) = n^2 + \sum_u \sum_v n^2_{uv} - \sum_u n^2_{u.} - \sum_v n^2_{.v}$$

$$c' = 2c = \sum_i \sum_{i'} \bar{c}^1_{ii'} c^2_{ii'} = \sum_i \sum_{i'} (1 - c^1_{ii'}) c^2_{ii'} = \sum_v n^2_{.v} - \sum_u \sum_v n^2_{uv}$$

$$d' = 2d = \sum_i \sum_{i'} c^1_{ii'} \bar{c}^2_{ii'} = \sum_u n^2_{u.} - \sum_u \sum_v n^2_{uv}$$

3.4 Indice de Rand

3.4.1 Indice de Rand Brut

Dans le but de comparer deux partitions de classes p et q, l'indice d'accord le plus utilisé est l'indice de Rand. Cet indice brut R de Rand (semblable au taux de Kendall pour les variables ordinales) est le pourcentage global de paires en accord *:*

$$R = \frac{A}{\binom{n}{2}}$$

On montre que :

$$A = \binom{n}{2} + \sum_u \sum_v n^2_{uv} - \frac{1}{2}\left[\sum_u n^2_{u.} + \sum_v n^2_{.v}\right]$$

En utilisant les notations du tableau Tab.3.1, l'indice de Rand peut être sous la forme suivante :

$$R=(a+d)/(a+b+c+d)$$

L'indice de Rand écrit sous sa forme contingentielle selon Marcotorchino [MAR 91] où on considère toutes les paires, y compris celles identiques est :

$$R' = \frac{2\sum_u \sum_v n^2_{uv} - \sum_u n^2_{u.} - \sum_v n^2_{.v} + n^2}{n^2}$$

Il prend ses valeurs entre 0 et 1 ; Il est égal à 1 lorsque les deux partitions sont identiques.

En utilisant les formules de linéarisation, N. EL Ayoubi [MAR 91] a montré que cette dernière version de R peut être écrite sous la forme relationnelle suivante :

$$R' = \frac{1}{n^2}\left[\sum_i \sum_{i'} c^{1}_{ii'} c^{2}_{ii'} + \sum_i \sum_{i'} c^{-1}_{ii'} c^{-2}_{ii'}\right]$$

C'est avec cette formulation relationnelle qu'Idrissi [IDR 00] a étudié la normalité asymptotique de R' sous l'hypothèse d'indépendance. A titre d'exemple, si les k classes (dans le cas où les deux partitions ont même nombre de classes p=q=k) sont équiprobables on trouve que $c^{1}_{ii'} c^{2}_{ii'} + c^{-1}_{ii'} c^{-2}_{ii'}$ suit une loi de Bernoulli de paramètre : $1 - \frac{2}{k} + \frac{2}{k^2}$, on en déduit :

$$E(R') = 1 - \frac{2}{k} + \frac{2}{k^2}$$

A. Idrissi affirme que le coefficient de Rand empirique entre deux variables qualitatives à k modalités équiprobables calculées sur n observations suit asymptotiquement une loi normale de variance :

$$V(R') = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{k} + \frac{2}{k^2}\right)\left(\frac{2}{k} - \frac{2}{k^2}\right)$$

Cette expression de la variance suppose l'indépendance des $c_{ii'}$, ce qui est inexact en raison des contraintes de transitivité ($c_{ik} = c_{ii'} . c_{i'k}$) et n'est vraie qu'approximativement pour k grand (il n'y a même pas normalité asymptotique pour des partitions en deux classes).

3.4.2 Indice de Rand corrigé selon Huber et Arabie

Pour deux partitions aléatoires, la valeur espérée de l'indice de Rand n'est pas nulle. L'indice de Rand ajusté proposé par [HUB 85] a pour forme générale :

$$\frac{\text{indice - indice espéré}}{\text{indice maximum - indice espéré}}$$

Cet indice qui peut être au plus égal à 1, prend donc la valeur 0 quand l'indice est égal à l'indice espéré. Avec une hypothèse de distribution hypergéométrique, on montre que l'indice de Rand corrigé R_C égal à :

$$R_C = \frac{R - R_{esp}}{R_{max} - R_{esp}} = \frac{n^2 \cdot \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}{\frac{1}{2} \cdot n^2 \cdot (\sum_u n_{u.}^2 + \sum_v n_{.v}^2) - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}$$

L'indice maximum R_{max} étant égal à 1, tandis que l'indice espéré R_{esp} s'obtient en remplaçant n_{uv} dans l'expression de R par $\frac{n_{u.} \cdot n_{.v}}{n}$. On peut noter qu'on aurait obtenu le même coefficient R_C, si on avait fait le calcul à partir de R'.

L'indice de Rand brut est souvent plus élevé que celui corrigé. Hubert et Arabie affirment que la correction augmente la sensibilité de cet indice. L'espérance de l'indice corrigé est nul lorsque les accords entre les deux partitions sont dus au hasard ; Cependant cet indice corrigé peut prendre des valeurs négatives lorsque les partitions sont peu liées.

3.4.3 Indice de Rand dans sa version asymétrique.

Définition

Dans le cas où on a deux partitions d'un même ensemble d'individus mais avec des nombres de classes inégaux, on utilise l'indice de Rand asymétrique proposé par [CHAV 01]. Cet indice asymétrique évalue dans quelle mesure une partition P_1(souvent experte) est « plus fine » qu'une partition P_2. Lorsque la partition experte est engendrée par une variable qualitative, on peut simplement vouloir qu'une classe de la partition obtenue contienne tous les objets d'une ou de plusieurs classes de la partition experte P_1. P_1 aura alors en général plus de classes que P_2 et il semble plus naturel d'utiliser des critères de comparaison non symétrique.

On considère les deux partitions P_1 et P_2 de n individus dont le nombre de classes de P_1 est supérieur au nombre de classes de P_2. P_1 est plus fine que P_2 lorsque deux éléments sont classés ensemble dans P_1 ils le sont également dans P_2 : $\forall$ u =1,…,p, $\exists$ v tel que $P^1_u \subseteq P^2_v$, P_u (respectivement P_v) désignant la $u^{ème}$ (respectivement $v^{ème}$) classe de P_1

(respectivement P_2). On cherche ainsi à mesurer l'inclusion de la partition P_1 dans la partition P_2.

Nous présentons une écriture simple où nous considérons toutes les paires y compris celles identiques. Ce critère de Rand asymétrique, noté R_A, est défini par :

$$R_A(P_1, P_2) = 1 + \frac{\sum_{u,v}\binom{n_{uv}}{2} - \sum_{u}\binom{n_{u.}}{2}}{\binom{n}{2}}$$

R_A prend ses valeurs dans l'intervalle [0,1]. Si $\forall u, \exists v$ tel que $P_u^1 \subseteq P_v^2$, alors $R_A=1$.

En considérant toutes les paires d'individus, y compris celles identiques on peut écrire cette version de R_A de la façon suivante :

$$R_{A'}(P_1,P_2) = \frac{n^2 + \sum_{u,v} n_{uv}^2 - \sum_{u} n_{u.}^2}{n^2} = \frac{a'+b'+c'}{a'+b+c'+d'}$$

Notons que dans le cas où les deux partitions auraient même nombre de classes, l'indice de Rand asymétrique n'est pas égal à l'indice de Rand brut.

Indice de Rand Asymétrique corrigé

Il est difficile de déterminer les cas où R et R_A sont nuls. Le critère de Rand asymétrique corrigé utilise aussi la normalisation $\frac{R_A - R_{Aesp}}{1 - R_{Aesp}}$ et vaut donc :

$$R_{AC} = \frac{n^2 \cdot \sum_{u,v} n_{uv}^2 - \sum_{u} n_{u.}^2 \cdot \sum_{v} n_{.v}^2}{n^2 \cdot \sum_{u} n_{u.}^2 - \sum_{u} n_{u.}^2 \cdot \sum_{v} n_{.v}^2}$$

En notant :

$N_{uv} = n_{uv}$ = nombre d'individus qui sont dans les classes u de P_1 et v de P_2

$N_{u.} = n_{u.}$ = nombre d'individus qui sont dans les classes u de P_1

$N_{u\bar{v}} = n_{u.} - n_{uv}$ = nombre d'individus qui sont dans la classe u de P_1 et ne sont pas dans la classe v de P_2

$N^{-}_{uv} = n_{.v} - n_{uv}$ = nombre d'individus qui sont dans la classe v de P_2 et ne sont pas dans la classe u de P_1

on obtient l'écriture suivante :

$$R_{A'}(P_1, P_2) = \frac{1}{n^2}\sum_u\sum_v N_{uv}^2 + \frac{1}{n^2}\sum_u\sum_v N_{u.}N^{-}_{uv}$$

qui peut être réécrite par les formules de comparaison par paires :

$$R_{A'}(P_1, P_2) = 1 - \frac{1}{n^2}\sum_i\sum_{i'} c^{1}_{ii'}c^{-2}_{ii'}$$

Sous l'hypothèse d'indépendance et d'équiprobabilité on trouve que l'espérance de $R_{A'}$ est :

$$E(R_{A'}) = 1 - \frac{1}{p} + \frac{1}{pq}$$

3.5 Un indice inspiré de Mc Nemar

Le test de Mc Nemar est un test non-paramétrique bien connu utilisé pour vérifier l'égalité de deux proportions dans des échantillons appariés (par exemple pourcentage d'individus favorables à une certaine opinion avant et après une campagne). Si a désigne le nombre d'individus qui ont gardé la même opinion favorable, avant et après, b le nombre d'individus qui ont gardé la même opinion défavorable, c et d les effectifs de ceux qui ont changé d'avis (Tab. 3.1), la statistique de test correspondant à l'hypothèse nulle H_0 selon laquelle les changements d'opinion dans un sens ou d'autre sont équiprobables est :

$$M_c = \frac{d - c}{\sqrt{d + c}}$$

Mc suit approximativement une loi normale N(0,1) sous H_0.

En adaptant le test de Mc Nemar à l'ensemble des paires d'individus, on a une nouvelle façon de mesurer la concordance entre deux partitions, qui revient à se demander si les paires qui sont séparées le sont par hasard, donc on étudie le désaccord entre les paires d'individus. On montre facilement que :

$$Mc = \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{2\sqrt{\frac{1}{2}(\sum_u n_{u.}^2 + \sum_v n_{.v}^2) - \sum_u \sum_v n_{uv}^2}} = \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{\sqrt{2}\sqrt{\sum_u n_{u.}^2 + \sum_v n_{.v}^2 - 2\sum_u \sum_v n_{uv}^2}}$$

On trouve la forme relationnelle de cet indice représentée de la façon suivante :

$$Mc = \frac{\sum_i \sum_{i'} c_{ii'}^1 - \sum_i \sum_{i'} c_{ii'}^2}{\sqrt{2(\sum_i \sum_{i'} c_{ii'}^1 \bar{c}_{ii'}^2 + \sum_i \sum_{i'} \bar{c}_{ii'}^1 c_{ii'}^2)}}$$

3.6 Indice de Jaccard

L'indice de Jaccard est un coefficient d'association connu pour étudier la similarité entre objets pour des données binaires de présence-absence.

Le tableau binaire suivant représente un exemple de présence-absence de deux individus i et i' quelconques à m critères différents :

	v_1	v_2	v_3	… v_m
i	1	1	0	1
i'	0	1	0	0

Tab. 3.2 *Tableau binaire de présence- absence*

On peut former alors le tableau suivant :

i / i'	1	0
1	$11_{(i,i')}$	$10_{(i,i')}$
0	$01_{(i,i')}$	$00_{(i,i')}$

T**ab. 3.3** *Tableau croisant les deux individus selon les m critères*

Où $11_{(i,i')}$ = nombre de critères ou propriétés que i et i' possèdent simultanément

$01_{(i,i')}$ = nombre de propriétés que i ne possède pas mais que i' possède

$10_{(i,i')}$ = nombre de propriétés que i' ne possède pas mais que i possède

$00_{(i,i')}$ = nombre de propriétés que i et i' ne possède pas .

l'indice J de Jaccard s'écrit de la façon suivante :

$$J(i,i') = \frac{11(i,i')}{11(i,i') + 10(i,i') + 01(i,i')}$$

Cet indice varie entre 0 à 1 et ne tient compte que des associations positives (présences simultanées). Par analogie, on définit l'indice de Jaccard d'accord entre deux partitions par :

$$J = \frac{a}{a + c + d}$$

on trouve alors que l'indice s'écrit :

$$J = \frac{\sum_u \sum_v n_{uv}^2 - n}{\sum_u n_{u.}^2 + \sum_v n_{.v}^2 - \sum_u \sum_v n_{uv}^2 - n}$$

En utilisant les formules de passages contingence-paires, on trouve la forme relationnelle de cet indice :

$$J = \frac{\sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} - n}{\sum_i \sum_{i'} c^{-1}_{ii'} c^2_{ii'} + \sum_i \sum_{i'} c^1_{ii'} - n}$$

3.7 Indice de corrélation vectoriel RV d'Escoufier

Toute classification repose sur les distances inter-individus : Pour trouver la ressemblance entre deux études faites sur les mêmes observations, on peut comparer les deux tableaux de distances inter-individus ou les matrices de produits scalaires associées. Le coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] permet de mesurer la similarité de deux configurations en tenant compte de la possibilité d'avoir différent métriques pour mesurer les distances entre les individus.

RV mesure la ressemblance entre deux tableaux de données numériques X_1 et X_2 sur les mêmes observations en comparant les produits scalaires inter-individus associés aux deux tableaux. Ces matrices de produits scalaires $X_iX'_i$ notées W_{ii} sont de dimensions nxn. Le coefficient RV est défini par :

$$RV(X_1, X_2) = \frac{\text{trace}(W_{12} W_{21})}{\sqrt{\text{trace}(W_{11}^2)\,\text{trace}(W_{22}^2)}}$$

Il est la somme des carrées inter- covariance entre les deux tableaux X_1 et X_2 divisé par la matrice normée intra-variance.

Les travaux de A. Lazraq et R.Cléroux [LAZ 01,02] donnent la possibilité de tester des hypothèses concernant RV mais pour des données numériques.

Si on applique ce coefficient à deux tableaux disjonctifs K_1 et K_2, on trouve :

$$RV(P_1,P_2) = \frac{\text{trace}\,(C^1C^2)}{\sqrt{\text{trace}\,(C^1)^2\,\text{trace}\,(C^2)^2}} = \frac{\sum_{i,i'}(c^1_{ii'})(c^2_{ii'})}{\sqrt{\sum_{i,i'}(c^1_{ii'})^2\sum_{i,i'}(c^1_{ii'})^2}} \qquad (*)$$

Si RV est suffisamment grand, les classifications obtenues seront voisines.

3.8 Indice JV de Janson et Vegelius

Une nouvelle façon de trouver la ressemblance entre deux partitions provenant d'un même ensemble de données, est d'utiliser le coefficient JV introduit par S. Janson et J. Vegelius [JAN 82]. Il correspond au critère de l'écart à la moyenne probabiliste CP sous l'hypothèse de l'équiprobabilité des classes pour une partition [IDR 00].

Le coefficient JV s'écrit sous sa forme relationnelle et contingentielle (pour p>2 et q>2)

$$JV(P_1,P_2)= \frac{\sum_{i,i'}(c^1_{ii'}-\frac{1}{p})(c^2_{ii'}-\frac{1}{q})}{\sqrt{\sum_{i,i'}(c^1_{ii'}-\frac{1}{p})^2\sum_{i,i'}(c^2_{ii'}-\frac{1}{q})^2}} = \frac{pq\sum_u\sum_v n^2_{uv} - p\sum_u n^2_{u.} - q\sum_v n^2_{.v} + n^2}{\sqrt{[p(p-2)\sum_u n^2_{u.} + n^2][q(q-2)\sum_v n^2_{.v} + n^2]}}$$

Idrissi [IDR 00] a utilisé cette formule pour étudier la distribution probabiliste de JV sous l'hypothèse d'indépendance. Dans le cas où les k modalités de deux variables qualitatives seraient équiprobables, on trouve que $\sum_{i\neq i'} c^k_{ii'}c^l_{ii'}$ suit une loi Binomiale de paramètres (n.(n-1), $\frac{1}{k^2}$) L'espérance et la variance de JV sous les même conditions sont égal à :

$$E\,(JV)= \frac{k-1}{n}$$

Si on centre les $c_{ii'}^{k}$ dans l'équation (*), le coefficient de corrélation vectoriel RV et l'indice de Janson et Vegelius JV s'identifient.

3.9 Indice de Redondance

L'indice de redondance proposé par Stewart et Love [STE 68] est un indice non symétrique et il est défini par :

$$RI(X_1,X_2)= \frac{\text{trace}(W_{12}W_{22}^{-1}W_{21})}{\text{trace}(W_{11})}$$

C'est une moyenne pondérée des carrées des coefficients de corrélation multiple entre les composantes de X_1 et X_2. Il sert à mesurer la qualité de prédiction de X_1 par X_2. Il est la proportion de variance expliquée dans la régression de X_1 par X_2. Il est utilisé, en autres, pour la sélection de variables en régression linéaire multivariée.

Dans le cas où X_1 et X_2 seraient les tableaux de variables indicatrices, [SAP 90] a montré que l'indice RI n'est autre que l'indice de dépendance non symétrique τ_b de Goodman et kruskal [GOO 79]. τ_b mesure le taux de décroissance du pourcentage de prédictions incorrectes. Pour deux partitions P_1 et P_2, τ_b s'écrit sous sa forme contingentielle de la façon suivante :

$$\tau_{bP_2/P_1} = \frac{\sum_u \sum_v \frac{n_{uv}^2}{n.n_{u.}} - \sum_v \left(\frac{n_{.v}}{n}\right)^2}{1 - \sum_v \left(\frac{n_{.v}}{n}\right)^2}$$

Par définition, $0 \leq \tau_b \leq 1$, avec τ_b=0 dans le cas de l'indépendance, et τ_b=1 pour la liaison fonctionnelle.

Si on utilise les formules de passages contingences-paires, τ_b peut être écrit sous la forme relationnelle suivante :

$$\tau_b = \frac{n.\sum_i \sum_{i'} \frac{c^1_{ii'} c^2_{ii'}}{c^1_{i.}} - \sum_i \sum_{i'} c^2_{ii'}}{n^2 - \sum_i \sum_{i'} c^2_{ii'}}$$

Il est connu que $0 \leq RI \leq 1$. RI devient égale à ρ^2, le carré du coefficient de corrélation simple lorsque p=q=1. RI se réduit au carré du coefficient de corrélation multiple lorsque p=1 et q>1.

R. Cléroux et al. [LAZ 02] a obtenu l'estimation S-robuste de l'indice de redondance et il a construit un test asymptotique robuste pour $H_0 : \rho I = 0$ contre $H_1 : \rho I > 0$ lorsque la loi sous–jacente est dans la classe des lois elliptiques.

Le test est de rejeter H_0 au niveau α si $n .RI > c_\alpha$ où c_α est la valeur critique qui peut être obtenue en utilisant l'algorithme exact de Imhof (1961) dont un programme Fortran est donnée dans Koerts et Abrahamse (1989).

Lorsqu'on utilise cet indice non symétrique pour comparer deux partitions de différents nombres de classes, tant que RI est assez grand, on peut constater que les deux partitions sont proches.

Notons que cet indice asymétrique est intéressent pour savoir si des classes sont apparues ou non dans deux instants différents d'enquêtes.

3.10 Coefficient Kappa de Cohen

Introduit par Cohen [COH 60], le test non paramétrique kappa mesure l'accord entre deux variables qualitatives pour des données appariées. Prenons le cas de deux praticiens examinant un même patient, et qui proposent des diagnostiques différents. Il est important que l'accord soit le meilleur possible pour garantir la qualité et la continuité des soins.

Définition

Kappa exprime une différence relative entre la proportion d'accord observée P_o et la proportion d'accord aléatoire P_e qui est la valeur espérée sous l'hypothèse nulle

d'indépendance des variables, divisée par la quantité disponible au-delà de l'accord aléatoire.

Ce coefficient est le pourcentage de l'accord maximum corrigé de ce qu'il serait sous le simple effet du hasard.

Dans le cas d'étude d'accord entre deux variables indépendantes ayant k modalités, le coefficient kappa s'écrits :

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

La concordance observée P_o est la proportion des individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille de l'échantillon n.

$$P_o = \frac{1}{n} \sum_{i=1}^{k} n_{ii}$$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisés par le carrée de la taille de l'échantillon.

$$P_e = \frac{1}{n^2} \sum_{i=1}^{k} n_{i.} n_{.i}$$

Le coefficient kappa est un nombre réel, sans dimension, compris entre –1 et 1. L'accord sera d'autant plus élevé que la valeur de kappa est proche de 1 et l'accord maximal est atteint lorsque P_o=1 et P_e=0.5.

Lorsqu'il y a indépendance entre les variables, le coefficient kappa est nul et dans le cas d'un désaccord total, kappa prend la valeur –1 avec P_o=0 et P_e=0.5. Ceci n'est vrai que dans le cas où les marginales seraient égales ($n_{i.} = n_{.i}$) puisqu'il suffit de prendre les effectifs diagonaux (ceux qui expriment l'accord dans le tableau de contingence) égaux aux marginales et les effectifs non diagonaux égaux à 0.

Pour des marginales données, [COH 60] propose de déterminer la valeur maximale de Kappa (κ_m) :

$$\kappa_m = \frac{P_m - P_e}{1 - P_e}$$

où P_m la proportion d'accord maximal est sous la forme suivant :

$$P_m = \frac{1}{n} \sum_{i=1}^{k} \inf(n_{i.}, n_{.i})$$

Kappa appliqué au tableau de contingence

En utilisant le coefficient kappa au tableau de contingence, on trouve une nouvelle façon pour mesurer l'accord entre deux partitions provenant d'un même échantillon, dans le cas où on pourrait numéroter les classes.

Le coefficient kappa mesure l'écart à la diagonale du tableau de contingence, et s'écrit :

$$\kappa = \frac{n. \sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.} n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.} n_{.i}}$$

Quand on compare deux partitions de même données et de même nombre de classes, l'identification des classes est nécessaire pour utiliser kappa.

Pour chercher les numéros de classes, on permute le tableau de contingence trouvé par une méthode de partition, ici les k-means, et à chaque permutation on calcule la valeur du coefficient kappa de Cohen, la permutation qui donne la valeur maximale de kappa indique la numérotation des classes recherchée.

3.11 Indice D_2 de Popping

Définition

L'indice D_2 proposé par R. Popping [PO 83, PO 00], est l'un des indices de similarités qu'on peut utiliser. Il est basé sur le même principe que le coefficient kappa.

L'indice D_2, basé sur la comparaison des paires d'individus, étudie l'agrément entre deux juges qui caractérisent indépendamment le même ensemble de données dans le cas où les catégories ne seraient pas connues à l'avance. IL mesure l'accord positif entre deux variables nominales. Cet indice contient une correction d'agrément qui peut être espérée par chance donnant les marginaux de la classification originale.

Sous l'hypothèse d'indépendance, l'indice D_2 est une transformation du coefficient de Russel et Rao qui a la forme $\frac{a}{a+b+c+d}$ (Tab. 3.1).

D_2 est défini par :

$$D_2=\frac{D_o - D_e}{D_m - D_e}$$

D_o est l'accord positif, D_e est la valeur espérée de D_o sous les conditions d'indépendances entre les variables. D_m est la valeur maximale que D_o peut prendre. Avec :

$$Do= \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(n_{ij}-1)}{n(n-1)}$$

$$De= \frac{2\sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}}{n(n-1)}$$

$$c_{ij} = g_{ij}(h_{ij} - 0{,}5g_{ij} - 0{,}5) \quad \text{avec} \quad h_{ij} = \frac{n_{i.}.n_{.j}}{n} \quad \text{et} \quad g_{ij} = \text{Entier}(h_{ij})$$

$$D_p=\frac{\sum_{i=1}^{p} n_{i.}(n_{i.}-1)}{n(n-1)} \quad , \quad D_q=\frac{\sum n_{.j}(n_{.j}-1)}{n(n-1)}$$

$$D_m=\text{Max}(D_p,D_q)$$

Dans D_2, on considère D_e comme étant un minimum raisonnable [POP 83], donc l'utilisation de g_{ij} est favorisée car on n'a pas une démonstration empirique pour avoir une moyenne plus petite que le minimum, pourtant elle donne une valeur biaisée de D_e.

Notant que Popping [POP 94] a proposé l'indice S_2 qui mesure l'accord global positif et négatif entre deux variables nominales.

Lien entre D2, k, JV et J

L'indice D2 a été comparé au coefficient kappa et à l'indice de JV, dans le cas particulier suivant :

Catégorie	+	-	Total
+	e	h-e	h
-	h-e	e	h
Total	h	h	2h

Tab. 3.4 *Cas particulier du tableau d'accord utilisé par Popping*

Popping a obtenu les résultats suivants : $\kappa = \frac{2e - h}{h}$

$$D_2 = \left[\frac{2e - h}{h}\right]^2 = JV$$

Généralement, dans notre étude par simulation, nous trouvons une forte corrélation entre les indices D_2 et JV.

Pour notre part, si on considère que la partie Entière(h_{ij}) = h_{ij} , nous obtenons la relation suivante entre D_2 et l'indice de Jaccard J :

$$D_2 = \frac{(C_n^2 - b)J}{\max(a + c, a + d) - C_n^2}$$

avec $J = \frac{a}{a + c + d}$ et $C_n^2 = \frac{n(n-1)}{2} = \binom{n}{2}$

3.12 Conclusion

Ce chapitre a été consacré à la présentation des indices utiles pour notre procédure de comparaison de deux partitions. Nous avons présenté l'indice de Rand sous sa forme brute et corrigée ainsi que sa version asymétrique, les indices dérivés du test de Mc Nemar et de l'indice de Jaccard, l'indice de corrélation vectorielle, l'indice JV de Janson et Vegelius, l'indice asymétrique de redondance, le coefficient Kappa de Cohen, et l'indice D_2 de Popping. Nous avons proposé pour la plupart de ces indices des écritures relationnelles et contingentielles. Nous avons montré que RV s'identifie à JV et proposé une relation entre D_2 et J.

Notons que, l'indice de Khi-deux n'a pas été retenu pour cette étude car il étudie l'écart à l'indépendance et non pas à la concordance. Ceci sera justifié dans le chapitre suivant qui traitera la comparaison de deux partitions provenant d'une même ensemble d'individus.

Chapitre 4

Comparaisons de deux partitions sur les mêmes individus

4.1 Introduction

Il est fréquent d'avoir à comparer des partitions provenant d'un même ensemble d'individus, obtenues dans diverses circonstances (opinion, consommation, méthodes, enquêtes, algorithme …), plusieurs cas se posent, selon que l'on travaille sur le même questionnaire ou sur des questionnaires différents. La littérature spécialisée est souvent muette sur ce point.

Dans ce chapitre, on propose des méthodes et des approches destinées à répondre aux différentes questions : Lors de deux enquêtes portant sur les mêmes individus, comment mesurer l'accord entre deux classifications ? Est- ce que ces deux classifications se ressemblent ? Peut-on affirmer que la classification n'a pas changé, que les proportions respectives des classes ont ou non pas varié, que les classes s'interprètent de la même façon ?

En mesurant la ressemblance entre deux classifications provenant d'un même ensemble d'individus, on distinguera trois cas:

- **Variables totalement différentes,** lorsque les deux classifications à comparer sont obtenues par différentes variables. A titre d'exemple, on peut comparer l'attitude vis à vis de deux produits bancaires pour les mêmes clients.
- **Variables partiellement différentes,** sont les variables communes (supplémentaires ou illustratives), qui ne contribuent pas à la construction des classes mais qui sont utilisées a posteriori pour identifier et caractériser les

regroupements établis à partir des variables actives. C'est l'exemple d'une comparaison de deux produits utilisés par les même clients, où on considère comme variables supplémentaires, les variables age, sexe, profession, etc.….

- **Variables identiques,** c'est le cas des données appariées ou panels, où un même échantillon d'individus est soumis à deux mesures successives des même variables.

4.2 Le Problème de la numérotation des classes

Quand on a à comparer deux partitions des mêmes individus, obtenues par des méthodes de classification, la numérotation des classes est arbitraire. L'identification des classes est indispensable pour l'utilisation de certain indice tel que le coefficient kappa de Cohen. Notre approche consiste à proposer une méthode pour identifier les classes de deux partitions provenant d'un même ensemble d'individus. La démarche se résume à choisir la permutation des numéros des classes de l'une des deux partitions dans le tableau croisé, qui donne la valeur maximale du coefficient kappa.

Les résultats de cette démarche seront confrontés avec ceux obtenus par d'autres méthodes tel que : la méthode au moyen de l'analyse factorielle de correspondances qui cherche à maximiser le poids de la diagonale du tableau de contingence croisant les deux partitions [SAP 90], [LEB 97], la méthode graphique de Bertin [BER 77] introduite dans le logiciel AMADO (Analyse graphique d'une Matrice de Données) [RIS 94], et la méthode d'analyse de données symboliques qui cherche à minimiser la distance entre les descriptions symboliques des classes en utilisant les variables supplémentaires.

Pour découvrir et identifier la structure d'un phénomène quantifié sous forme d'un tableau croisé en analyse de données, nous disposons de deux approches :

- La technique numérique en analyse de correspondance, permettent de découvrir rapidement les grands traits de la structure du tableau.
- La Graphique (terme créé par Jacques Bertin) utilise l'extraordinaire outil d'analyse qu'est l'œil pour découvrir et montrer les ressemblances et oppositions entre éléments (lignes ou colonnes) du tableau.

4.2.1 Méthode par maximisation du kappa

L'identification des classes est nécessaire pour utiliser le coefficient kappa κ : il est alors logique d'identifier les classes des partitions qui conduisent à une valeur maximale de κ. On renumérote les classes de l'une des deux partitions du tableau de contingence pour optimiser κ. On prend alors la permutation des classes qui maximise κ. En effet, pour chercher les numéros de classes, on permute les colonnes ou lignes du tableau de contingence croisant les deux partitions, et à chaque permutation on calcule la valeur de κ. La permutation qui donne la valeur maximale de kappa indique la numérotation des classes recherchée.

Considérons deux partitions de 1000 individus à quatre classes. Le tableau de contingence de base de l'une des itérations croisant P_1 (en ligne) et P_2 (en colonne) est le suivant :

P_1 \ P_2	**1**	**2**	**3**	**4**
1	248	0	0	2
2	1	198	27	9
3	2	6	43	202
4	0	58	192	12

Tab. 4.1 *Tableau de contingence initial de P_1 et P_2*

Ce tableau donne une valeur du coefficient Kappa égale à 0.33506. Afin d'identifier les classes de P_2 à celles de P_1, on réordonne les colonnes pour obtenir la valeur de kappa maximale (il y a 4 ! permutations). Le tableau réordonné est alors le suivant:

P_1 \ P_2	**1**	**2**	**4**	**3**
1	248	0	2	0
2	1	198	9	27
3	2	6	202	43
4	0	58	12	192

Tab. 4.2 *Tableau de contingence réordonné selon la numérotation du κ maximal*

La valeur maximale de kappa est de 0.786698 pour la permutation suivante des colonnes : 1,2,4,3. La classe 3 de P_1 est alors identifiée à la classe 4 de P_2 et vice-versa.

En utilisant la formulation de κ_m proposée par [COH 60] qui tient compte des effectifs marginaux, on obtient une valeur égale à 0.96267. Le rapport de cette dernière avec la valeur obtenue par notre méthode est :

$$\frac{0\,.786698}{0\,.96267} \times 100 \quad \approx \quad 81\%$$

Ce rapport montre que l'accord obtenu par notre méthode correspond à 81% de l'accord maximal qu'il pourrait atteindre.

4.2.2 L'Analyse Factorielle des Correspondances

Pour trouver l'ordre optimum des classes, on peut utiliser l'Analyse Factorielle des Correspondances (AFC). Cette méthode offre la possibilité d'interpréter et de positionner un point d'un ensemble relatif à un espace par rapport à l'ensemble des autres points définis dans l'autre espace. Les nuages de points-lignes et de points-colonnes vont être représentées dans les plans de projection formés par les premiers axes factoriels pris deux à deux.

En AFC, qui est une analyse canonique entre deux groupes d'indicatrices, le premier axe a la propriété suivante: les coordonnées des catégories des variables sont les valeurs numériques telles que leur coefficient de corrélation linéaire soit maximal. Il est donc logique de permuter les modalités selon leur classement sur cet axe.

On réordonne lignes et colonnes du tableau de contingence selon l'ordre des points sur le premier axe principal pour trouver un tableau dont les termes « diagonaux » aient des effectifs maximaux.

En appliquant cette méthode aux deux partitions P_1 et P_2 du paragraphe précédent, et par utilisation du logiciel SPAD (méthode de CORBI), le tableau de contingence réordonné est:

P_1 \ P_2	**2**	**3**	**4**	**1**
2	198	27	9	1
4	58	192	12	0
3	6	43	202	2
1	0	0	2	248

Tab. 4.3 *Ordre selon le premier facteur en AFC*

On trouve la même correspondance entre classe qu'en maximisant kappa.

4.2.3 Méthode graphique de Bertin

La graphique [BER 77] propose de « faire voir » le tableau croisé par la juxtaposition d'histogrammes (ou graphiques « en tuyaux d'orgues »). Cette représentation améliore considérablement la lisibilité des résultats de l'analyse statistique des données multidimensionnelles [RIS 94].

La graphique est complémentaire à l'analyse de correspondances pour découvrir rapidement la structure du tableau, structure diagonale qui sera lisible.

Le reclassement visuel des lignes et des colonnes avec la graphique permet de compléter et d'affiner les classements automatiques (AFC, ACP, pourcentage, tri) et le plus souvent de les déplacer en classant « à part » lignes et colonnes appartenant à des systèmes différents. Ce reclassement aboutit à une image qui dégage clairement une structure évolutive et des exceptions facilement analysables.

Le logiciel AMADO [RIS 94] permet de réaliser ces graphiques et donne la possibilité de réaliser facilement des améliorations du graphique par déplacement de ligne ou/et colonnes. On peut corriger l'ordre, en partie aléatoire, des éléments de l'arbre de classifications (dendrogramme) et améliorer la lecture des résultats en utilisant les informations non inscrites dans le tableau croisé.

En appliquant la méthode Graphique à la simulation des deux partitions P_1 et P_2, par utilisation du logiciel AMADO qui se trouve dans le logiciel SPAD, on trouve le graphique suivant :

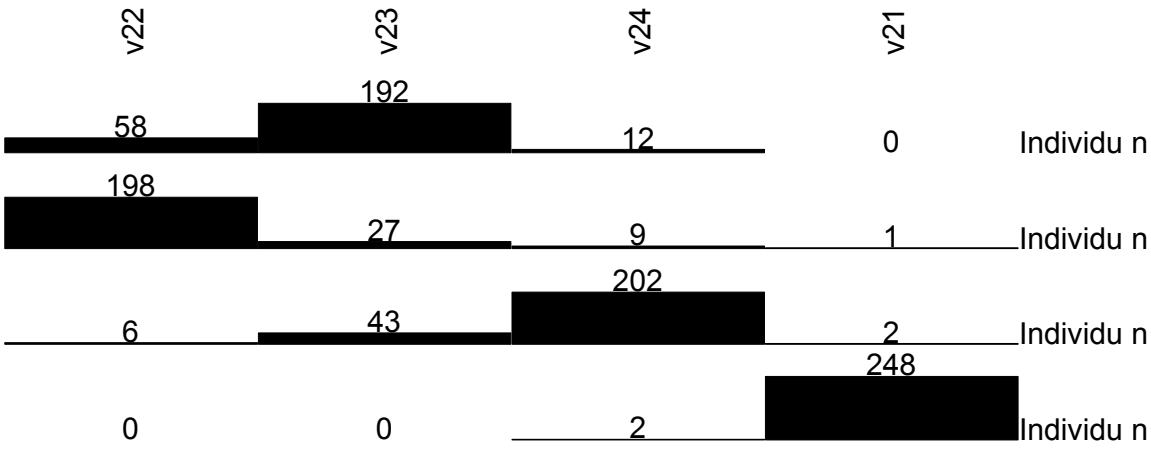


Fig. 4.1 *Graphique trouvé par la méthode de Bertin*

On retrouve de nouveau la même identification des classes que précédemment.

4.2.4 L'Analyse Symbolique

L'analyse des données symboliques offre la possibilité de caractériser chaque classe d'une partition par des propriétés qui sont les descriptions symboliques. Une idée pour identifier ces classes est de comparer les descriptions des classes par l'utilisation des mesures de similarités.

Notre approche consiste à minimiser la distance de dissimilarités entre les descriptions symboliques des classes en basant sur les variables illustratives. On groupe les objets symboliques de faible dissimilarités dans une classe homogène. Plusieurs indices de dissimilarités peuvent être utilisés. On présente dans la suite les différents indices de ressemblances (dissimilarités ou de similarités) pour comparer des objets symboliques.

Indices de Ressemblance

u et v deux objets symboliques définis par:

$u= [Y_1 \epsilon U_1]^ [Y_2 \epsilon U_2]^\ldots\ldots^[Y_p \epsilon U_p]$ et $v= [Y_1 \epsilon V_1]^ [Y_2 \epsilon V_2]^\ldots\ldots^[Y_p \epsilon V_p]$

Plusieurs mesures de dissimilarités sont proposées pour des objets symboliques booléens. On présente celles de Gowda et Diday [GOW 94], l'approche par Ichino et Yaguchi [ICH 94], et celui de Decarvalho [DEC 94, 98].

Gowda et Diday (1991) :

$$D(u,v)=\sum_{j} D(U_j, V_j)$$

- Pour des variables réelles, la distance entre U_j et V_j est :

$$D(U_j,V_j)=D_p(U_j,V_j)+D_s(U_j,V_j)+D_c(U_j,V_j)$$

Où D_p est le composant du à la position, D_s celui du à l'étendue ou le cardinal de deux descriptions, et D_c celui qui compare le contenu de deux descriptions.

- Pour des variables nominales ou ordinales, la distance est :

$$D(U_j,V_j)= D_s(U_j,V_j)+D_c(U_j,V_j)$$

Ichino et Yaguchi (1994) :

Pour une paire (U_j, V_j) liant à la j'ieme variable Yj, la distance est :

$$\phi(U_j, V_j) = \left|U_j \oplus V_j\right| - \left|U_j \cap V_j\right| + \gamma(2\left\|U_j \cap V_j\right\| - \left|U_j\right| - \left|V_j\right|)$$

pour $0 \le \gamma \le 0.5$. La distance de type Minkowski est alors:

$$d_q(a,b) = (\sum_{j=1}^{p} \phi(A_j, B_j)^q)^{1/q}$$

De Carvalho (1994) :

Les différentes fonctions pour une paire (U_j, V_j) de la j^{eme} variable Y_j ,en utilisant la terminologie du Tab 3.1:

$$d_1([Y_j \epsilon U_j],[Y_j \epsilon V_j])=1-\frac{a}{a+c+d}$$

$$d_2([Y_j \epsilon U_j],[Y_j \epsilon V_j])=1-\frac{2a}{2a+c+d}$$

$$d_3([Y_j \epsilon U_j],[Y_j \epsilon V_j])=1-\frac{a}{a+2(c+d)}$$

$$d_4([Y_j \epsilon U_j],[Y_j \epsilon V_j])=1-\frac{1}{2}\left[\frac{a}{a+c}+\frac{a}{a+d}\right]$$

$$d_5([Y_j \epsilon U_j],[Y_j \epsilon V_j]) = 1 - \frac{a}{\sqrt{(a+c)(a+d)}}$$

ces fonctions se transforment par les fonctions de distance entre les objets assertion u et v par les écritures suivantes:

$$d_q^i(u,v) = \left(\sum_{j=1}^{p} [w_j d_i(U_j, V_j)]^p \right)^{1/q}$$

Pour tout i $\epsilon\{1,2\ldots.,5\}$ quand w_j est le poids de la variable Y_j. Pour une paire (U_k,V_k) de valeur de la k ^{ième} variable, la distance est définie par :

$$\Psi'(U_j, V_j) = \frac{\Phi(U_j, V_j)}{\upsilon(U_j \oplus V_j)} \quad \text{tel que } \upsilon(U_j) = \begin{cases} |U_j| & \text{si } Y_j \text{ est nominale, ou ordinale} \\ |\overline{u}_j - \underline{u_j}| & \text{si } Y_j \text{ est continue} \end{cases}$$

et la correspondance entre deux objets u et v est alors :

$$d_q'(u,v) = \left(\sum_{j=1}^{p} \left[\frac{1}{p} \Psi'(U_j, V_j) \right]^p \right)^{1/q}$$

De Carvalho (1996,1998)

Pour deux objets d'assertion u et v, les fonctions de distance proposées sont :

$$d_1'(u,v) = \pi(u \oplus v) - \pi(u \otimes v) + \gamma(2.\pi(u \otimes v) - \pi(u) - \pi(v)) \qquad \text{avec} \qquad \pi(u) = \prod_{j=1}^{p} \upsilon(U_j) \text{ et}$$

$\otimes$ représente l'intersection.

$$d_2'(u,v) = \frac{\pi(u \oplus v) - \pi(u \otimes v) + \gamma(2.\pi(u \otimes v) - \pi(u) - \pi(v))}{\pi(u^E)}$$

pour $u^E = [x_1 = Y_1] \wedge \ldots [x_p = Y_p]$

$$d_3'(u,v) = \frac{\pi(u \oplus v) - \pi(u \otimes v) + \gamma(2.\pi(u \otimes v) - \pi(u) - \pi(v))}{\pi(u \oplus v)}$$

De Carvalho a utilisé la définition des descriptions potentielles pour prend on considération la dépendance logique entre les variables, l'extension de la distance de Minkowski devient :

$$d_q''(u,v) = \left(\frac{\sum_{j=1}^{p} \left| [d_i(U_j, V_j)]^q \right|}{\sum_{j=1}^{p} \delta(j)} \right)^{1/q}$$

avec NA est la valeur prie dans le cas où la dépendance hiearchique serait active

$$\text{et } \delta(j) = \begin{cases} \{NA\} & \text{si } Y_j \text{ est une conclusion des regles exprimant une dependance logique} \\ \phi & \text{ailleurs} \end{cases}$$

Pour une liste complète des extensions des coefficients d'associations, voir [ESP 00]. La comparaison entre ces mesures de dissimilarités pour des objets symboliques booléennes a été étudié par [MAL 01].

4.3 Démarche pour comparer deux partitions « proches »

4.3.1 Algorithme

Dans le but de comparer deux partitions provenant d'un même ensemble de données, nous avons proposé une méthodologie [SAP 02], [YOU 03, 04] qui consiste à étudier la distribution des indices d'associations en engendrant par simulation des partitions « proches » afin de donner des valeurs critiques sous des hypothèses réalistes.

Il faut maintenant définir ce que l'on entend par « deux partitions sont proches » : notre approche consiste à dire que les individus proviennent d'une même partition commune, dont les deux partitions observées en sont des réalisations bruitées. Le modèle de classes latentes est bien adapté à cette problématique pour engendrer des partitions. Notons qu'il a été utilisé récemment pour la recherche de partitions consensus par Green et Krieger [GRE 99].

Pour obtenir des partitions « proches », qui ne diffèrent l'une de l'autre que de façon aléatoire, on va construire des échantillons artificiels issus d'un modèle à k classes latentes et décrits par v variables numériques, que l'on supposera par commodité,

normales, mais d'autres distributions sont bien sûr possibles. On partage ensuite arbitrairement les v variables en deux groupes et on effectue deux partitions en k classes des n individus selon ces deux groupes de variables à l'aide d'une méthode classique (les k-means ou nuées dynamiques.) Normalement, ces deux partitions doivent être peu différentes, on calcule les indices présentés dans le chapitre 3, on obtient un échantillon de valeurs de ces indices, sous l'hypothèse de « partitions proches » en itérant m fois, ce qui permet d'étudier leur distribution.

L'algorithme se déroule de la façon suivante pour une des m itérations:

- Tirage des effectifs des classes latentes selon une loi multinomiale M(n; π_1 , ..,π_k)
- Pour chaque classe, tirage de v variables normales indépendantes (c'est l'hypothèse fondamentale de l'indépendance locale du modèle des classes latentes).
- Calcul d'une partition P_1 sur v_1 variables et d'une partition P_2 sur les autres v-v_1 variables.
- Calcul des indices de comparaison de partitions.

En itérant le procédé, on obtient par simulation la distribution d'échantillonnage des indices d'associations. Ces configurations sont reproduites à chaque jeu de simulations, pour des paramètres différents des lois normales

4.3.2 Etude distributionnelle des indices de similarité

Nous avons appliqué la procédure précédente en nous limitant à 4 classes latentes équiprobables, 1000 individus, et 4 variables.

Les paramètres des distributions normales ont été choisis de telle sorte que pour chaque variable x_j, la valeur absolue de la différence entre les moyennes de la distribution normale de deux classes différentes soit plus grande d'une fois et demie de son écart-type :

$$| m_{kj} - m_{k'j}| > 1.5\ \sigma_j \qquad \forall\ j = 1, 2, 3,4 \ \text{ et } \forall\ k \text{ et } k' = 1, 2,3,4$$

m_{kj} et $m_{k'j}$ étant les moyennes respectives de la variable x_j dans les classes k et k', et σ_j l'écart-type de x_j .

Nous présentons dans la suite les résultats de nos simulations au nombre d'itérations m égal à 1000 et selon plusieurs choix de paramètres effectués avec le logiciel S-Plus [SPL 00]. Les indices choisis, pour comparer les deux partitions, sont l'indice de Rand dans toutes ses versions, l'indice dérivé de Mc Nemar, l'indice de Jaccard, le coefficient de Janson et Vegelius, le coefficient D_2 de Popping et le coefficient kappa.

Premier choix

Pour chaque classe, le premier choix de paramètres des 4 variables normales indépendantes est présenté dans le tableau suivant :

Classe n°1		***Classe n°2***		***Classe n°3***		***Classe n°4***	
X1	N(1.2,1.5)	X1	N(-2,1.5)	X1	N(5,1.5)	X1	N(8,1.5)
X2	N(-10,2.5)	X2	N(0,2.5)	X2	N(-17,2.5)	X2	N(3.8,2.5)
X3	N(6,3.5)	X3	N(12,3.5)	X3	N(13,3.5)	X3	N(-5,3.5)
X4	N(-20,4.5)	X4	N(-12,4.5)	X4	N(0,4.5)	X4	N(7,4.5)

Tab. 4.4 *Les distributions par classe du premier choix*

Pour une des 1000 itérations, la figure suivante donne la répartition spatiale dans le plan des deux premières composantes principales de l'une des deux partitions :

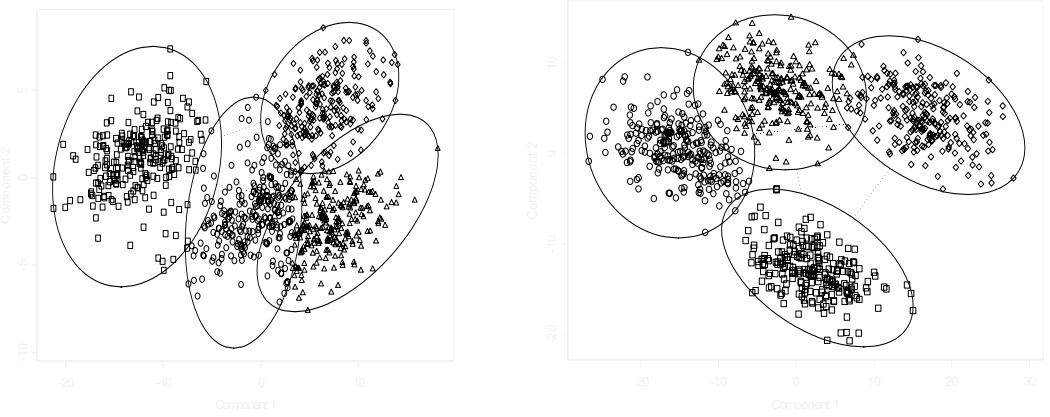


Fig. 4.2 *Les deux premières composantes principales de l'une des 1000 échantillons de P_1 et P_2 en premier choix*

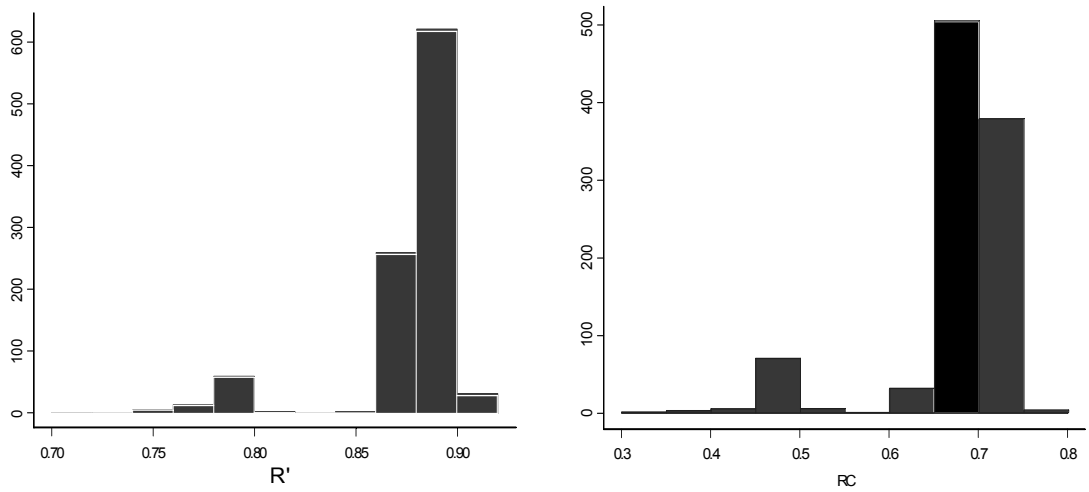


Fig. 4.3 *Distribution de l'indice de Rand(R') et de Rand Corrigé(RC)*

On remarque que la distribution de rand R' varie de 0.74 à 0.95 et celle de l'indice de RC varie de 0.3 à 0.8. Les deux indices ont une même allure de distribution. Mais n'oublions pas que l'indice de Rand cas corrigé peut avoir des valeurs négatives.

Ici, toutes les valeurs observées du coefficient de Rand sont supérieures à 0.7, alors que l'espérance de R' sous l'hypothèse d'indépendance est de 0.625, ce qui montre bien le caractère inadapté de celle-ci. Avec 1000 observations, on rejetterait l'indépendance si R'>0.65 au risque de 5% mais cela ne suffit pas pour montrer que les deux partitions sont « proches ».

Nous représentons dans la figure suivante la densité de la différence (R'- RC) de ces deux indices.

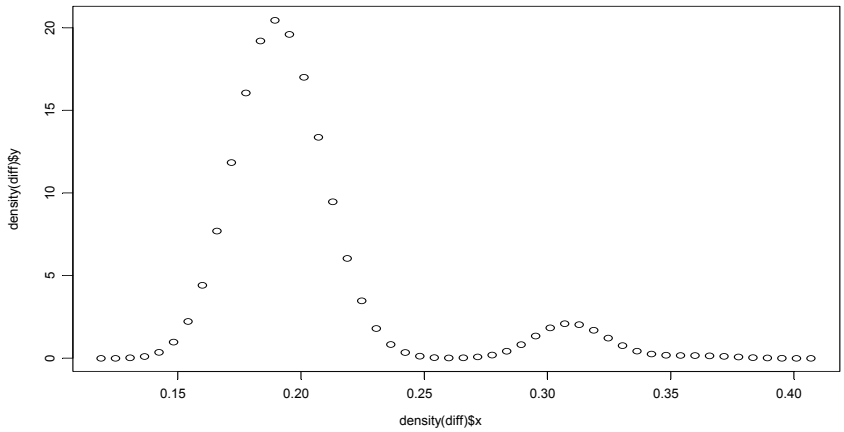


Fig. 4.4 *Densité de la différence entre l'indice de Rand R' et celui corrigé*

La différence entre l'indice de Rand R' et celui corrigé par Hubert [HUB 85] est très proche d'une distribution normale de moyenne de 0.19.

Comme l'indice de Rand donne la même importance aux couples d'individus qui sont dans la même classes de deux partitions, qu'à ceux qui ne sont pas dans la même classe pour les deux partitions (accord négatif), on utilise la même démarche pour trouver la distribution des indices de Mc Nemar et celui de Jaccard pour 1000 simulations.

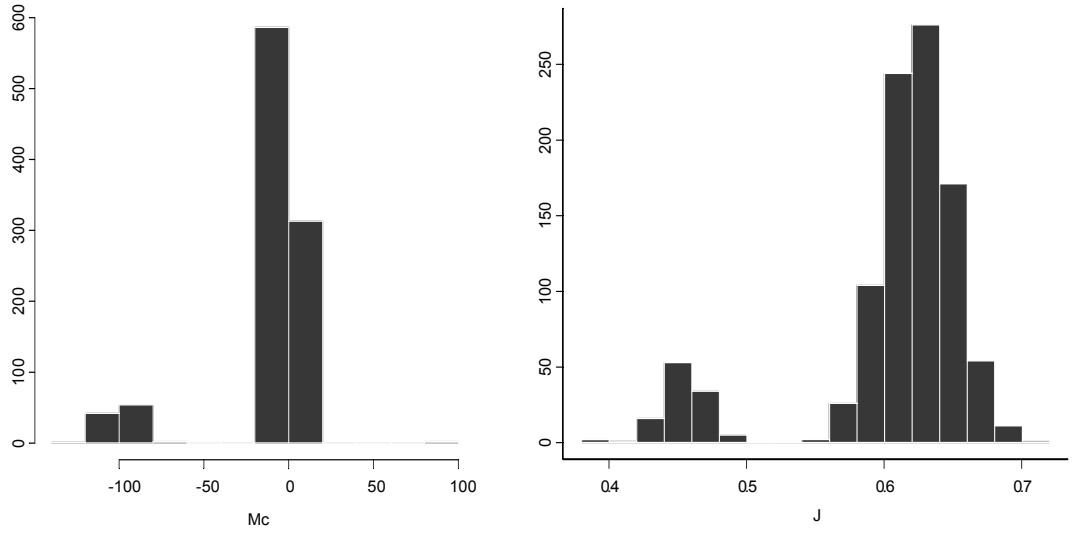


Fig. 4.5 *Distribution de l'indice de Mc Nemar et de Jaccard*

L'indice de Mc Nemar est à majorité distribué autour de zéro montrant ainsi que pour un risque de 5% l'hypothèse nulle est vérifiée. La distribution de l'indice de Jaccard présente des valeurs supérieures à 0.4 dont la valeur la plus fréquente est de 0.63.

On cherche les distributions de l'indice JV (ou RV) de Janson et Vegelius et l'indice D_2 proposé par Popping. La valeur de l'indice JV (Fig.4.6) varie entre 0.4 et 0.8, on trouve une distribution bimodale, et la valeur moyenne de l'indice JV est de 0.6712795. Les valeurs de l'indice D_2 varient entre 0.3 et 0.8. la valeur la plus fréquente est de 0.687 et la moyenne de la distribution est de 0.65.

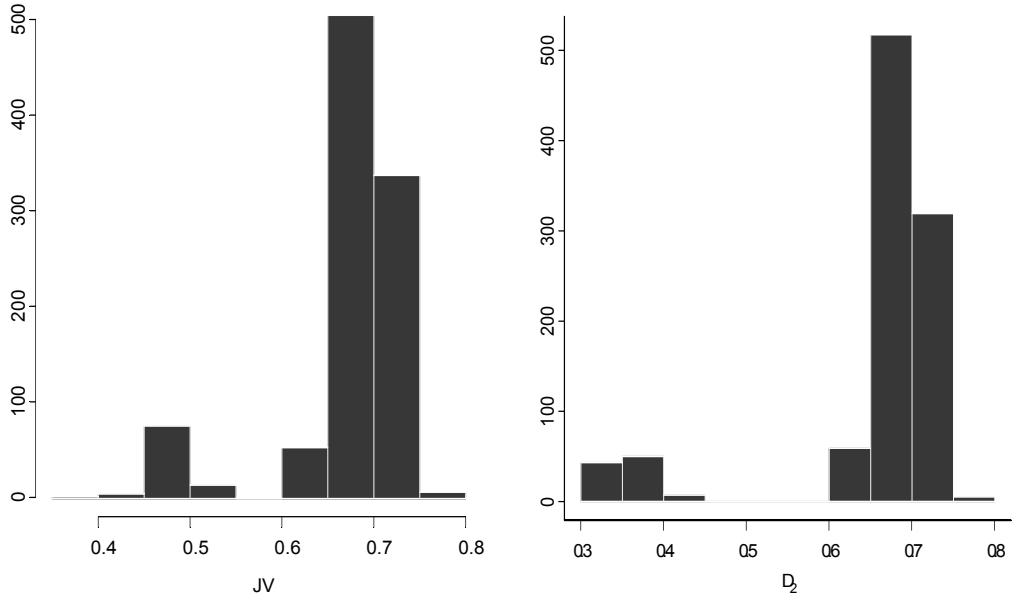


Fig. 4.6 *La distribution de JV et de D_2 pour le premier choix de paramètres*

Pour trouver la distribution de Rand asymétrique, la même procédure est utilisée pour trouver les variables normales indépendantes, mais en effectuant deux autres classifications : la première partition P_1 de X_1, X_2 et X_3 formée de 6 classes, et la deuxième partition P_2 de X_4 formée de 3 classes par K-means.

Dans ce cas, on cherche à évaluer dans quelle mesure les classes de P_1 sont incluses dans celles de P_2. On calcule alors 1000 fois l'indice de Rand asymétrique RA' et celui

corrigé. On remarque que les valeurs de l'indice de Rand asymétrique RA' sont supérieur à 0.92. Par contre celui de Rand asymétrique corrigé prend ses valeurs à partir de 0.36.

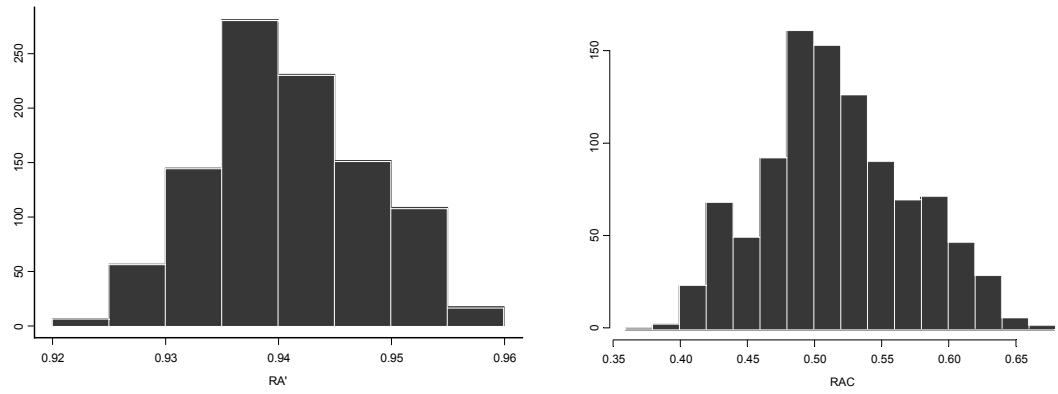


Fig.4.7 *Distribution de l'indice de Rand asymétrique et celui corrigé*

La représentation graphique de la densité de la différence entre l'indice de Rand asymétrique et celui corrigé (RA'-RAc) est donnée dans la figure suivante :

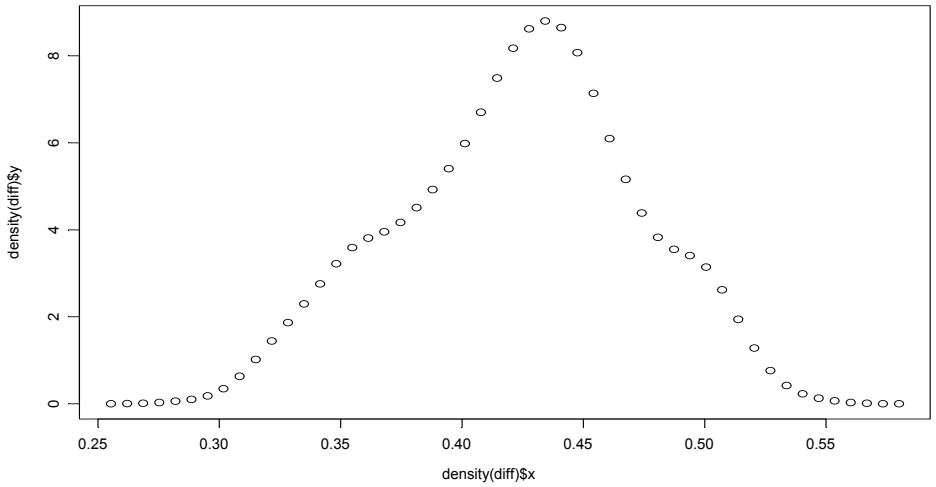


Fig.4.8 *Répartition de la densité de la différence de l'indicede Rand asymétrique RA' et celui corrigé* RA_C

La distribution de la différence de RA' et RA_C n'est pas une distribution normale, elle est de moyenne 0.435.

Afin de comparer l'indice de Rand brut R' et celui de Rand asymétrique RA', on représente la distribution de Rand brut R' pour cette même partition :

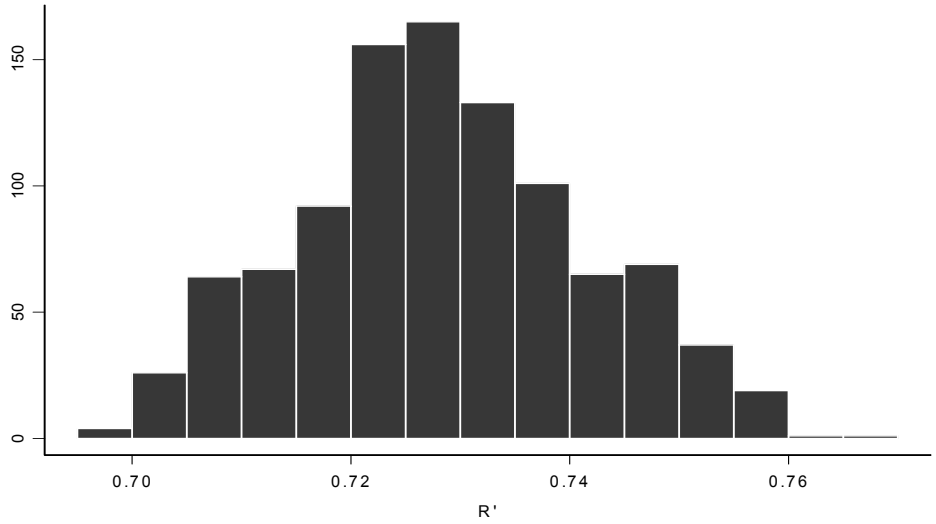


Fig.4.9 *Distribution de l'indice de Rand R'des partitions asymétriques*

Contrairement à ce qu'on a trouvé dans les résultats des partitions symétriques, on a une distribution modale dans tous les cas de l'indice de Rand. Cela revient à conclure que ces distributions dépendent du nombre de classes dans chaque partition.

On ne peut cependant proposer de seuil de signification pour chacun des coefficients, car les distributions dépendent de leur séparabilité qui est liée aux paramètres des distributions normales. On choisit un autre choix de paramètres des variables normales afin de tester leurs influences aux différents indices.

4.1.1.1 Deuxième choix de paramètres

Pour le deuxième choix, on choisit le modèle de mélange présenté dans le tableau suivant :

Classe n°1		***Classe n°2***		***Classe n°3***		***Classe n°4***	
X1	N(1.2,1.5)	X1	N(-2,1.5)	X1	N(-5,1.5)	X1	N(-8,1.5)
X2	N(4,2.5)	X2	N(-4,2.5)	X2	N(-10,2.5)	X2	N(-15,2.5)
X3	N(7,3.5)	X3	N(-6,3.5)	X3	N(-13,3.5)	X3	N(-20,3.5)
X4	N(10,4.5)	X4	N(-10,4.5)	X4	N(-20,4.5)	X4	N(-30,4.5)

Tab.4.5 *Le modèle de mélange normal choisi*

La figure suivante montre la répartition spatiale de l'une des 1000 itérations des deux partitions P_1 et P_2.

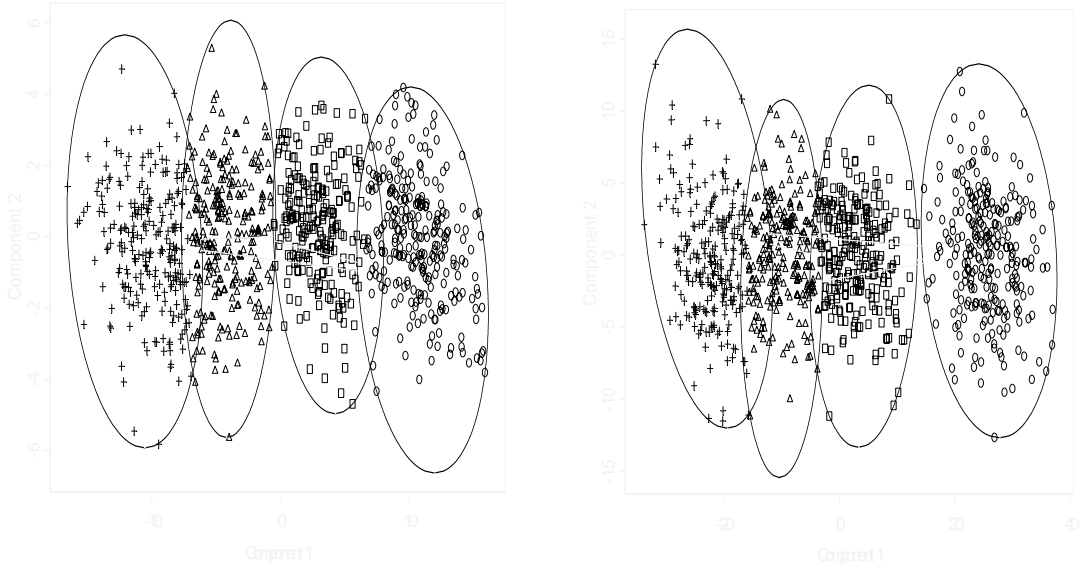


Fig. 4.10 *Les deux premières composantes principales de l'une des 1000 échantillons de P_1 et P_2 en deuxième choix*

Les distributions de l'indice R' de Rand brut, de l'indice dérivé de Jaccard et celui de Mac Nemar, sont représentées dans la figure suivante :

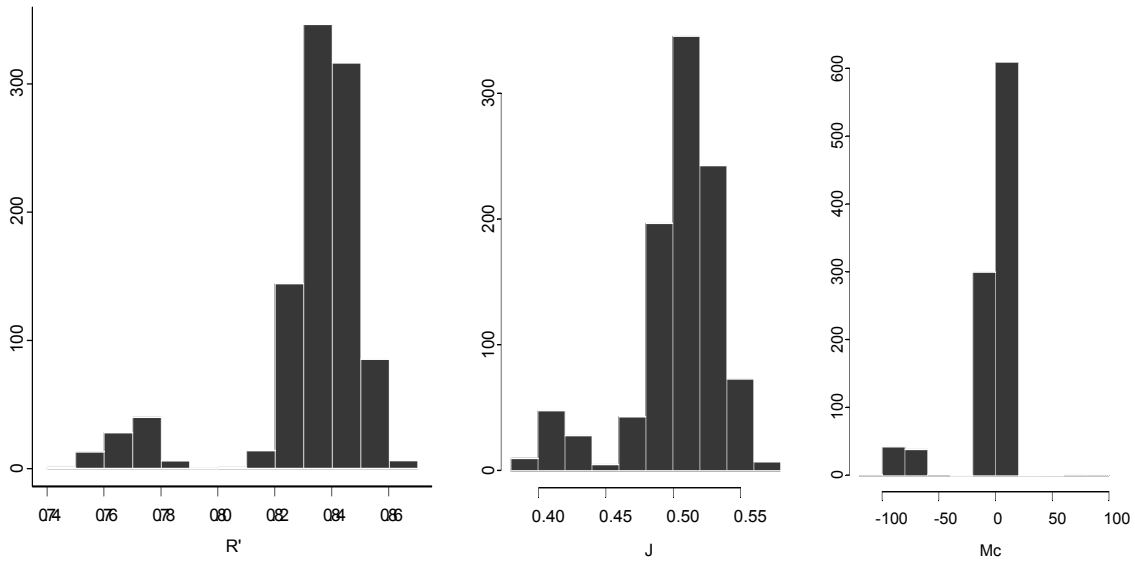


Fig. 4.11 *La distribution de R', J et de Mc pour le deuxième choix de paramètres*

On trouve une distribution de Rand R' toujours supérieure à 0.74. Cette distribution non normale varie entre 0.75 et 0.87 avec une moyenne égale à 0.8324716. La valeur la plus fréquente est de 0.835. L'indice de Jaccard prend ses valeurs entre 0.35 et 0.58. On observe une chute de ses valeurs par rapport à sa distribution précédente (Fig.4.5). La moyenne a baissé de 10% et prend la valeur de 0.5039546. L'indice Mc a toujours ses valeurs autour de zéro, sa distribution n'a pas donc changé d'allure.

Les corrélations entre R' et Mc, et entre J et Mc sont de valeurs respectives 0.8695076 et 0.7766866 qui sont faible par rapport à la corrélation entre R' et J dont la valeur est de 0.9807.

Pour visualiser l'allure de ces corrélations, on présente les nuages des points de ces indices deux à deux :

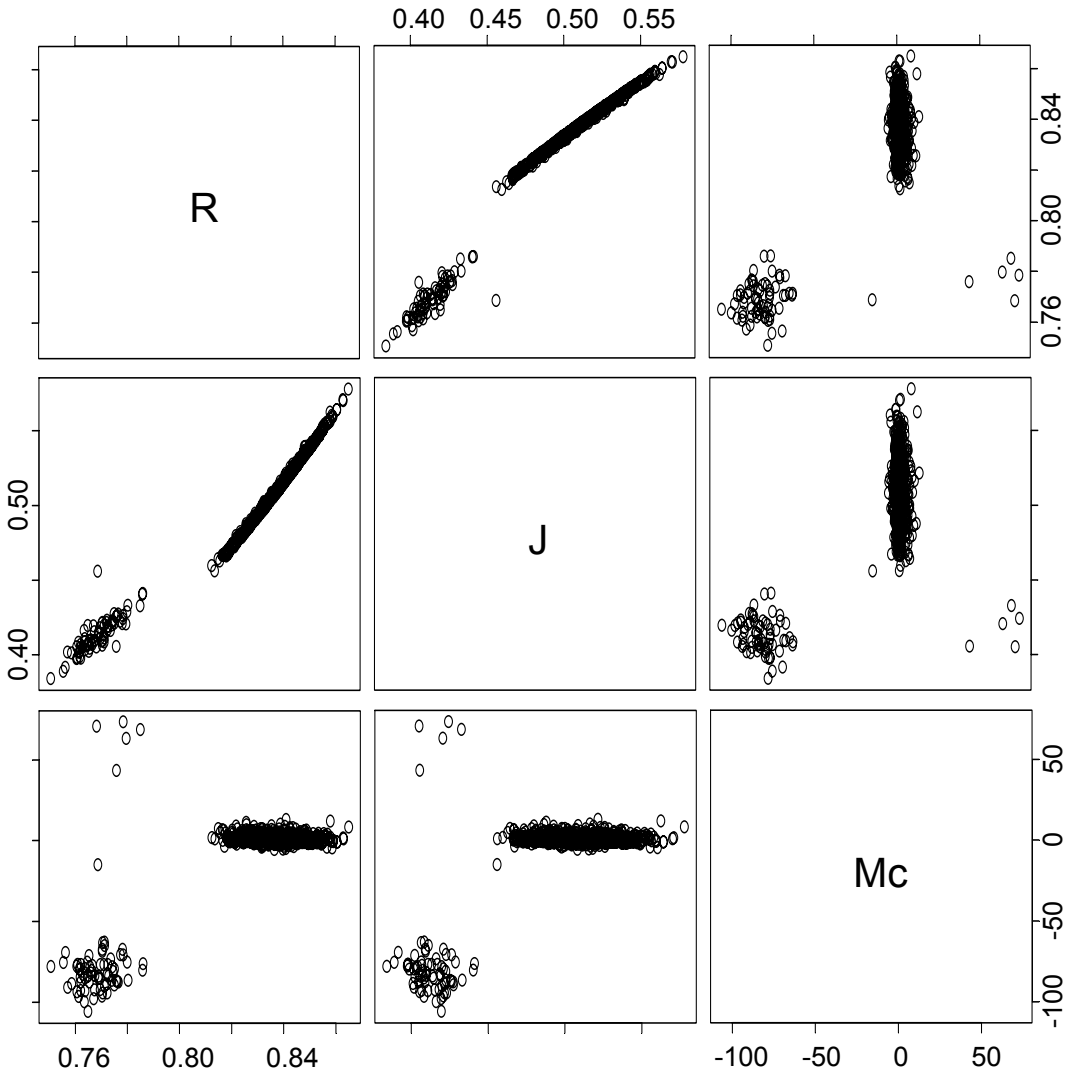


Fig. 4.12 *Nuage des points des indices R', J, et Mc l'un contre l'autre dans les 1000 itérations.*

On remarque la forte corrélation entre les deux indices R' et J, (Fig. 4.12) et les non-corrélation de l'indice Mc avec R' et J. La partie séparée de nuage des points des indices l'un contre l'autre provient de l'utilisation de la méthode des k-means qui donne un optimum local.

Pour ce deuxième choix, les résultats illustrés par la Fig. 4.13 montre que l'indice de JV prend ses valeurs entre 0.4 et 0.7. La valeur la plus fréquente est de 0.63 et la moyenne des valeurs est égale à 0.617.

Sous l'hypothèse d'indépendance E(JV)=0.003, donc avec les 1000 observations on rejette l'indépendance pour JV> 0.617 au seuil de confiance de 5%. La valeur critique trouvée au risque de 5% est beaucoup plus élevée que celle trouvée dans le cas d'indépendance. Cela montre encore une fois que la non-indépendance n'entraîne nullement la forte concordance entre deux partitions.

L'indice D_2 prend ses valeurs entre 0.3 et 0.7 avec un mode égal à 0.625. La distribution admet une moyenne de 0.6100 (Fig. 4.13).

Sous l'hypothèse nulle de partitions proches, toutes les partitions ont des valeurs autour de la moyenne dont la valeur est de 0.6, d'où on peut conclure que les deux partitions sont proches.

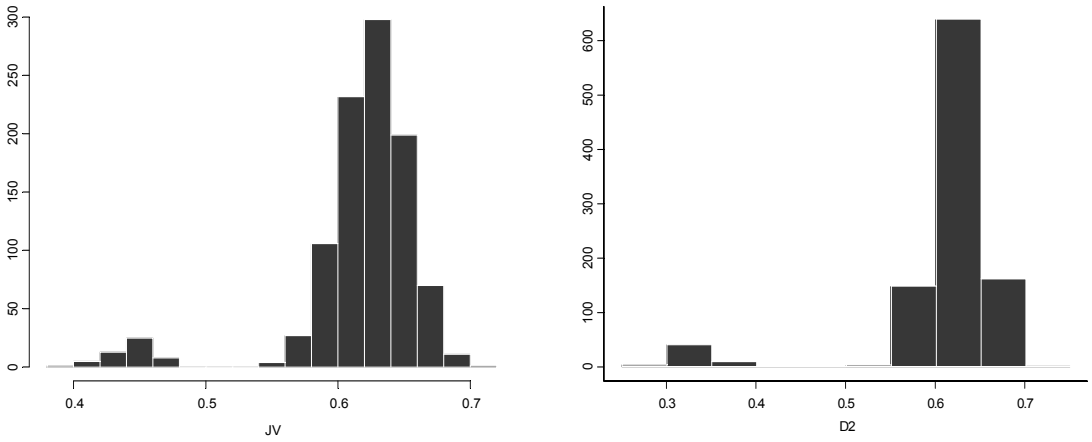


Fig. 4.13 *La distribution de JV et de D_2 pour des partitions à 4 classes en 1000 itérations*

On trouve une forte corrélation entre JV et D2 de valeur égale à 0.9832144. Nous pouvons donc conclure que l'utilisation de ces deux indices nous mène aux mêmes résultats en comparant deux partitions proches (Fig. 4.14).

La distribution du coefficient des valeurs de kappa maximal est présentée ci-dessous:

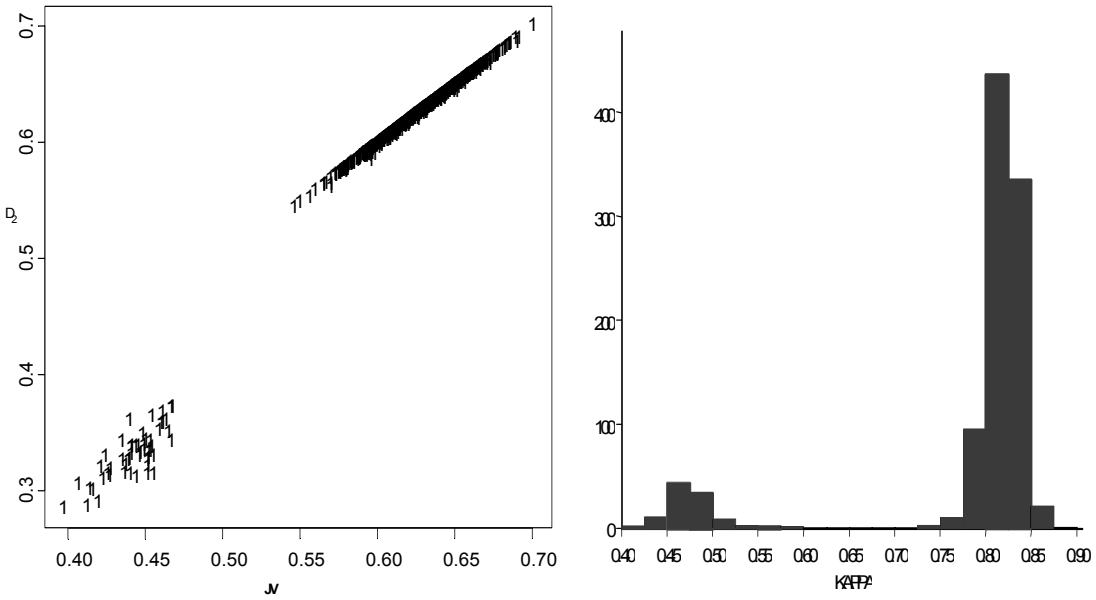


Fig. 4.14 *Distribution de kappa pour 1000 individus en 1000 itérations pour partitions à 4 classes. Nuage des points de JV contre D_2 dans les 1000 itérations.*

Avec ce choix de paramètres des variables normales indépendantes, on trouve que la distribution du coefficient de kappa varie entre 0.4 et 0.875. Sa moyenne est égale à 0.82.

En comparant les résultats des deux choix de paramètres, on remarque que la bimodalité est présente dans les deux cas causée par l'utilisation des k-means qui donne l'optimum locale [YOU 03]. L''indice de Mc n'a pas changé, l'indice de Jaccard a baissé de valeurs, les indices R', JV et D_2 ont à peu près les même moyennes dans les deux choix.

Il est clair que ces distributions dépendent de la plus ou moins grande séparation des classes, du nombre d'individus, et du nombre de classes des partitions. Ceci impose des simulations supplémentaires en faisant varier ces paramètres pour pouvoir conclure sur le bon choix d'indices.

4.1.1.2 Variation du nombre d'individus

En faisant varier le nombre d'individus des deux partitions, on obtient les résultats suivants :

n	**Moyenne de R'**	**Moyenne de J**	**Moyenne de Mc**	**Moyenne de JV**	**Moyenne de D_2**
40	0.9997277	0.5334094	-0.03614265	0.997844	0.6402844
200	0.9940175	0.5143862	-0.4955052	0.9557456	0.6933753
500	0.963623	0.5519325	-2.190844	0.8062435	0.6863509
1000	0.8562098	0.5039546	-5.904364	0.617	0.6100
1500	0.6232644	0.504464	-7.5382	0.459457	0.1896256

Tab.4.6 *La moyenne des indices par variation du nombre d'individus n en 1000 itérations.*

On remarque que la variation de ces indices n'est pas la même en fonction du nombre d'individus n. L'indice Mc décroît considérablement avec la variation de n, alors que l'indice J reste presque inchangé avec n.

4.1.1.3 Variation du nombre de classes des deux partitions

Pour 1000 itérations, et pour un type de choix de paramètres, on cherche la moyenne des différents indices en faisant varier le nombre de classes k de 3 à 8. Le résultat de cette procédure donne le tableau suivant :

N	n	k	Moyenne de R'	Moyenne de J	Moyenne de Mc	Moyenne de JV	Moyenne de D_2
1000	1000	3	0.8602054	0.6567402	-5.099624	0.4747902	0.4456527
1000	1000	4	0.8562098	0.5556003	-5.903291	0.617	0.6100
1000	1000	5	0.8916331	0.585066	0.0709304	0.6754467	0.6659623
1000	1000	6	0.8957413	0.5389056	7.577	0.6387779	0.617788
1000	1000	7	0.8990732	0.4858192	1.38892	0.5967879	0.5866291
1000	1000	8	0.8946015	0.4187404	2.347066	0.5325563	0.5203702

Tabl.**4.7** *Moyennes des indices par variation du nombre de classes k en 1000 itérations*

La moyenne de l'indice dérivé de Mc Nemar, de Janson et Vegelius JV et de Popping D_2 présente une variation qui passe par un maximum en fonction du nombre de classes latentes équiprobables. L'indice de Jaccard J décroît avec le nombre de classes. Cela peut s'expliquer en augmentant le nombre de classes, les paires d'individus (i, i') qui sont dans une même classe dans la première partition ont peu de chance de rester ensemble dans la deuxième partition. L'indice de Rand R' qui est l'indice le plus utilisé pour la comparaison de partitions, reste presque inchangé.

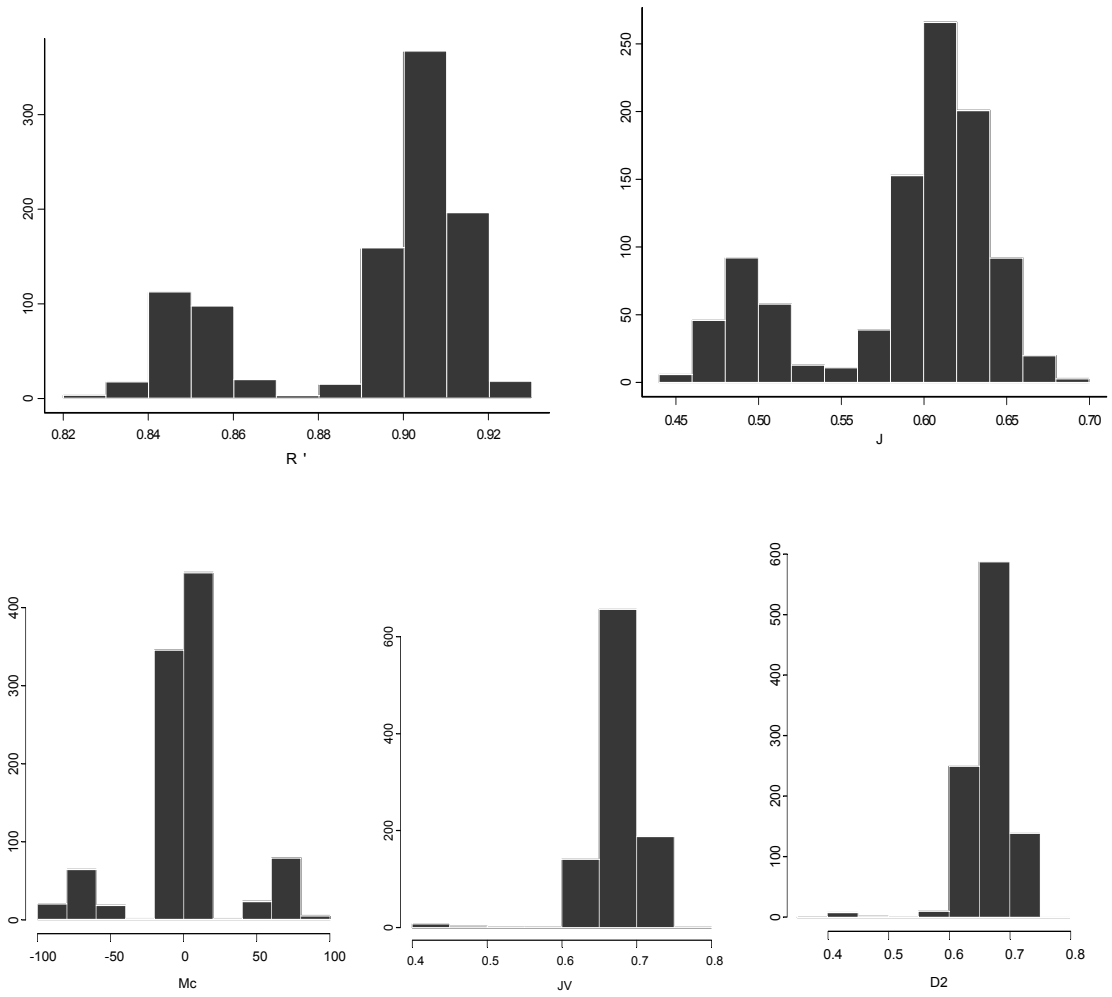


Fig. 4.15 *Distribution de l'indice de R', de J, de Mc, de JV et de D_2 cas de 5 classes*

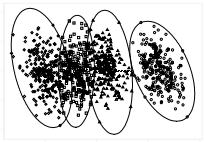
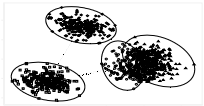
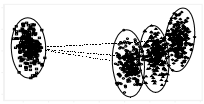
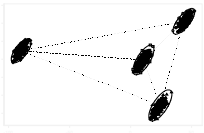
Pour un cas choisi du nombre de classe (k=5), on trouve la bimodalité (Fig.4.15) de la distribution de Rand à partir de 0.86. Les valeurs observées de l'indice de Jaccard sont

supérieures à 0.45. Les valeurs les plus fréquentes de l'indice de Mc Nemar sont à zéro. Les valeurs de JV et de D2 sont en majorités supérieures à 0.6.

4.1.1.4 Variation de la séparation des classes

Nous avons vu précédemment que pour deux choix distincts des paramètres de variables normales indépendantes, comment les indices proposés pour la comparaison de deux partitions proches varient selon la plus ou moins grande séparation des classes.

Afin d'avoir une idée sur la robustesse et la stabilité de ces indices selon la séparation des classes des partitions, on présente dans le tableau suivant les variations de ces indices pour différents types de choix des paramètres.

Type de séparation de classes	**Moyenne de R'**	**Moyenne de J**	**Moyenne de Mc**	**Moyenne de JV**	**Moyenne de D_2**
	0.832379	0.5033083	-3.834521	0.5580	0.548002
	0.83285	0.5373436	16.18574	0.584271	0.572841
	0.8774415	0.6230967	23.45929	0.686474	0.6570819
	0.9799209	0.9230288	-0.08404	0.946608	0.9460112
	0.9902108	0.9616261	-0.009685	0.9739486	0.9736587

Tabl.**4.8** *Moyennes des indices par variation de la séparation des classes en 1000 itérations*

L'indice de Rand R' est inchangé en fonction de la variation de paramètres des variables normales du modèle de mélange qui favorise la séparation des classes. Les indices JV, D_2 et J croient avec l'accroissement de la séparation des classes. Ces indices sont proches de 1, tant que les classes sont bien séparées.

Parmi ces indices, seul l'indice de Rand et de Janson et Vegelius possède, sous l'hypothèse d'indépendance, d'équiprobabilité, et du même nombre de classes, admettent une loi de distribution de probabilités théoriques [IDR 00]. Il serait donc intéressant de comparer leurs espérances théoriques de probabilités et celles simulées par notre méthode.

4.1.1.5 Comparaison entre la moyenne théorique et expérimentée de R' et de JV

- ***Indice Rand R'***

L'espérance théorique de l'indice de Rand sous l'hypothèse d'indépendance d'équiprobabilité et dans le cas de même nombre de classes est donnée par l'expression suivante : $E(R') = 1 - \frac{2}{k} + \frac{2}{k^2}$ [IDR 00]

En faisant varier le nombre de classes k des deux partitions, la simulation obtenue par utilisation de notre algorithme donne pour les 1000 itérations, les moyennes de l'indice de Rand R'. Pour cette même variation de k, on compare dans le tableau suivant ces moyennes avec leurs moyennes théoriques :

n	**k**	**E(R')**	**M_{exp} de R'**	$\Delta = \lvert M_{exp} - E(R') \rvert$
1000	3	0.55	0.8602054	0.3102054
1000	4	0.625	0.8562098	0.2312098
1000	5	0.68	0.8916331	0.2116331
1000	6	0.72	0.8833446	0.1633446
1000	7	0.775	0.8990732	0.1240732
1000	8	0.78125	0.8802063	0.0989563

Tab. 4.9 *Moyenne théorique et expérimentale de R' par variation de nombre de classe k en 1000 itérations*

La moyenne de Rand théorique croit avec le nombre de classes, mais ce n'est pas toujours le cas pour la moyenne expérimentale trouvée par simulation. La différence entre les valeurs théoriques et expérimentales décroît lorsque le nombre de classes des partitions augmente.

- ***Indice de JV***

L'espérance théorique de l'indice JV sous l'hypothèse d'indépendance, d'équiprobabilité et dans le cas de même nombre de classe selon Idrissi [IDR 00] est de :

$$E(JV)= \frac{k-1}{n}$$

En procédant à la même variation du paragraphe précédent, on obtient pour les 1000 itérations les moyennes suivantes :

n	**k**	**E(JV)**	**M_{exp} de JV**	$\Delta = \left\lvert M_{exp} - E(JV) \right\rvert$
1000	3	0.002	0.4747902	0.4727902
1000	4	0.003	0.617	0.614
1000	5	0.004	0.6754467	0.6714467
1000	6	0.005	0.6387779	0.6337779
1000	7	0.006	0.5967879	0.5907879
1000	8	0.007	0.360505	0.353505

Tab. 4.10 *Moyenne théorique et expérimentée de JV par variation de nombre de classe k en 1000 itérations*

On remarque une nette différence entre les deux moyennes théoriques et celles simulées. Cette différence diminue mais reste significatif avec l'accroissement du nombre de classes des deux partitions.

Alors les distributions de probabilités des indices de R' et JV proposées sous l'hypothèse d'indépendance par [IDR 00] ne sont évidement pas pertinés pour la question de la comparaison des partitions.

4.1.1.6 Sur la bimodalité

On a pu remarquer le caractère bimodal de la distribution du coefficient de Rand brut R' avec un mode secondaire correspondant à environ 10% des cas lorsque la séparation des classes n'est pas grande. Ce phénomène peut en fait s'expliquer par le caractère non-optimal de la méthode de classification utilisée : On sait que les k- means fournissent une solution dépendant du choix initial des centres. Or la procédure utilisée dans S+ fait une classification hiérarchique ascendante avant de lancer les k-means et part toujours de la même initialisation obtenue à partir d'une coupure en k classes de l'arbre hiérarchique. En modifiant le choix initial des centres par la procédure FASTCLUS de SAS nous avons pu obtenir des partitions finales différentes (souvent meilleures au sens de l'inertie) et une augmentation de l'indice de Rand. En appliquant une deuxième fois la méthode des k-means aux 10 % de valeurs correspondant au mode secondaire, on a pu obtenir une augmentation sensible de R'.

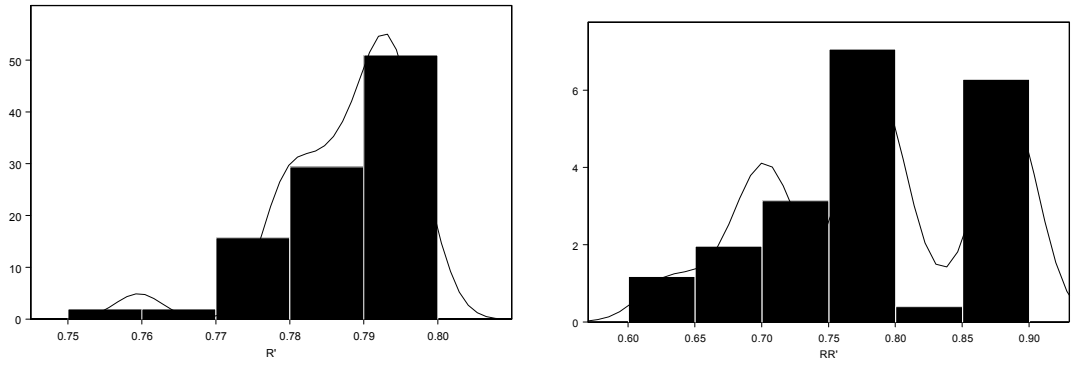


Fig. 4.16 *Distribution de l'indice de Rand R' en appliquant une et deux fois les k-means*

4.1.1.7 Discussion sur les indices

Dans le but de comprendre l'application des indices pour la comparaison de deux partitions « proches », nous venons de présenter quelques valeurs de ces indices obtenues par simulations. Ceci nous aident à développer quelques critères pour leur utilisation, rendant ainsi le choix moins « aléatoire ».

Sur la figure 4.17, on constate que l'indice de Rand R' reste presque inchangé et indépendant d'une part du nombre de classes k et d'autre part de la variation de

paramètres des variables normales du modèle de mélange. Il est presque stable en fonction de la séparation des classes. L'accroissement de nombres d'individus n provoque une décroissance de cet indice, remarquable pour les grands échantillons supérieurs à 1000 individus. L'utilisation de cet indice sera préférable pour la comparaison de deux partitions proches pour sa robustesse devant ces variations.

L'indice dérivé J de Jaccard est indépendant de la variation de nombres d'individus, d'où l'importance de son utilisation en deuxième lieu, mais il décroît avec l'accroissement du nombre de classes k.

L'indice D_2 forme une parabole de valeur maximale atteinte pour un nombre de classes k égal à 5 et il est stable pour un nombre d'individus inférieur à 1000.

L'indice JV a la même allure que D2 en fonction du nombre de classes, mais il décroît avec l'accroissement du nombre d'individus.

Avec l'accroissement de la séparation des classes, les indices JV, D_2 et J croient de la même manière (Fig.4.17).

L'indice Mc dépendant et trop variant avec tous les paramètres évoqués précédemment, ne présente aucun intérêt pour l'utilisation dans la comparaison de deux partitions proches et donc à déconseiller.

L'indice kappa de Cohen est utilisé seulement pour deux partitions de mêmes nombres de classes à condition d'identifier les classes par permutation. Son utilisation est donc plus complexe.

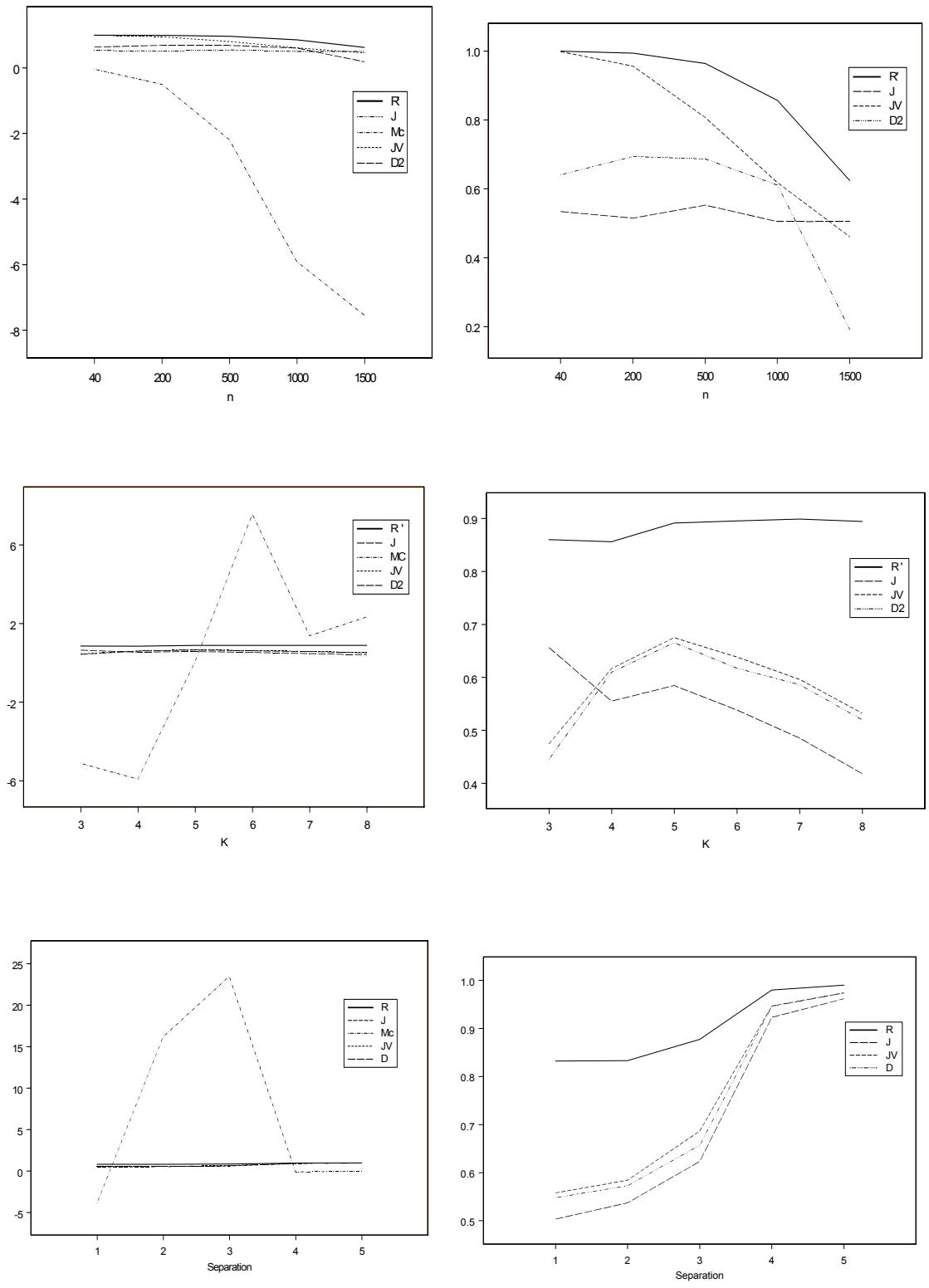


Fig. 4.17 *Allure de la variation de n, de k et de la séparation des classes pour R', J, Mc JV et* D_2

On remarque de ces graphiques que lorsque la séparation des classes de deux partitions est grande les indices se stabilisent à des valeurs très proches de 1.

Un point important pour comparer des partitions c'est de savoir si les classes à comparer sont stables et homogènes. On présente dans la suite les méthodes pour étudier la stabilité des classes des partitions.

4.4 Stabilités des classes

Dans cette partie, on s'intéresse à étudier la stabilité d'une classe d'une partition dans le but de répondre à la question suivante : Les classes des deux partitions signifient-elles le même, les proportions des classes ont–elles changés ? Surtout que certains indices demandent la nécessité d'identifier les numéros des classes avant qu'on les utilise comme l'indice kappa.

Nous présentons dans la suite quelques outils des tests classiques pour étudier la stabilité d'une classe de deux partitions d'une même base de données.

4.4.1 Test d'homogénéité de χ^2

On présente le test d'homogénéité qui s'applique dans le but est de savoir si les classes de deux partitions sont homogènes ou non, on teste alors l'hypothèse suivante:

H_0 : les partitions proviennent de la même population contre

H_1 : Les partitions sont significativement différentes.

Les k classes des deux partitions d'un même nombre n d'individus sont réparties de la façon suivante :

	$\mathbf{c_1}$	$\mathbf{c_2}$	….	$\mathbf{c_k}$	**Total**
$\mathbf{P_1}$	n_{11}	n_{12}		n_{1k}	n
$\mathbf{P_2}$	n_{21}	n_{22}		n_{2k}	n

Tab. 4.11 *Tableau de la répartition des classes selon P_1 et P_2*

Si n_{lh} est le nombre des individus de la partition P_l qui se trouve dans la classe h.

On a :

$n = \sum_{h=1}^{K} n_{1h} = \sum_{h=1}^{K} n_{2h}$ = la taille de la population pour les deux partitions.

$n_{.h} = n_{1h} + n_{2h}$ = nombre total des individus qui se trouve dans la classe r pour les deux partitions.

Pour l'hypothèse H_0, les $p_1, p_2, \ldots, p_k$ représentent les probabilités d'être dans les classes $c_1, c_2, \ldots, c_k$. Il s'agit donc de comparer les effectifs constatés n_{1h} ou n_{2h} aux effectifs espérés np_h qui ne doivent pas en différer beaucoup.

On a :

$$d^2 = \sum_h \frac{(n_{1h} - nph)^2}{nph} + \sum_h \frac{(n_{2h} - nph)^2}{np_h}$$

d^2 est une réalisation de D^2 suivant un χ^2 dont le degré de liberté est : 2k-2=2(k-1) (avec k est le nombre de classes dans une partition). On estime les $\hat{p}_h = \frac{n_{.h}}{2n} = \frac{n_{1h} + n_{2h}}{2n}$, ce qui fait (k-1) estimations indépendantes.

D'où

$$d^2 = \sum_h \frac{n_{1h} - \frac{n \times n_{.h}}{2n}}{\frac{n \times n_{.h}}{2n}} + \sum_h \frac{n_{2h} - \frac{n \times n_{.h}}{2n}}{\frac{n \times n_{.h}}{2n}} = \sum_h \frac{n_{1h} - \frac{n_{.h}}{2}}{\frac{n_{.h}}{2}} + \sum_h \frac{n_{2h} - \frac{n_{.h}}{2}}{\frac{n_{.h}}{2}}$$

D^2 est un χ^2 de degré de liberté 2k-2-(k-1) = χ^2_{k-1}, pour deux partitions homogènes. L'inconvénient est qu'il est un test sur les distributions marginales, il n'utilise pas le fait que les individus n'aient pas changé de classes.

4.4.2 Test de Mc Nemar

Pour étudier la stabilité des classes, on teste si les proportions des classes de deux partitions ont changé. Ce test non paramétrique mesure, pour des données dichotomiques, l'égalité des proportions des classes.

La statistique de test correspondant à l'hypothèse nulle H_0 selon laquelle les changements d'opinion dans un sens ou d'autre sont équiprobables est :

$$M_c = \frac{d - c}{\sqrt{d + c}}$$

Mc suit approximativement une loi normale N(0,1) sous H_0 (voir chapitre 3).

On utilise le test généralisé de Mc Nemar [GIL 89] qui étudie la variation des pourcentages sur un ensemble d'individus pour des classes des deux partitions. Il est utilisé pour tester si la probabilité d'individus classées dans (i, j) est la même que la probabilité d'individus classées dans (j, i). Pour deux partitions P_1 et P_2 formées de k classes chacune, le tableau de contingence est représenté comme suit :

P_1 \ P_2	**Classe 1**	**Classe 2**	**Classe v**	**totaux**
Classe 1	n_{11}	n_{12}	..	$n_{1.}$
Classe 2	n_{21}	n_{22}	..	$n_{2.}$
Classe u	..	…	n_{uv}	$n_{u.}$
totaux	$n_{.1}$	$n_{.2}$	$n_{.v}$	n

Tab. 4.12 *Tableau de contingence de P_1 et P_2*

Avec n_{uv}= nombre d'individus qui sont dans la classe u de P_1 et dans la classe v de P_2

L'hypothèse nulle est donnée par :

$H_0 : n_{u.} = n_{.u} \ \forall u \in k$ contre

$H_1 : \exists u' \text{ tel que } n_{u'.} \neq n_{.u'}$

La statistique du test de Mc Nemar dans le cas de notre tableau est écrit alors :

$$T = \sum_{u \neq v} \frac{(n_{uv} - n_{vu})^2}{n_{uv} + n_{vu}}$$

Elle peut être corrigée par la formule suivante :

$$T' = \sum_{u \neq v} \frac{(|n_{uv} - n_{vu}| - 1)^2}{n_{uv} + n_{vu}}$$

On rejette H_0 au niveau de confiance α si T dépasse le quantile $(1-\alpha)$ de la loi de khi-2 de degré de liberté égal à k(k-1)/2 où k est le nombre de classes de deux partitions. Sinon, on accepte à α% que les proportions des classes n'ont pas changé dans les deux partitions. Ce test a un avantage sur le test précédent car il tient en compte du fait que ce sont les mêmes individus. On va voir dans le chapitre 6 comment utiliser ce test par un exemple d'application.

4.5 Approches symboliques

4.5.1 Stabilité des classes d'objets symboliques

Pour étudier les qualités d'une classe pour des données symboliques, une façon simple consiste à associer à chaque classe un objet symbolique (en prenant l'union ou l'intersection des objets de la classe) puis à utiliser les propriétés et qualités des objets symboliques. Les caractéristiques de la classe sont : la stabilité et l'effritement.

Stabilité d'une classe

C'est la capacité d'une classe à être représenter par l'objet symbolique de plus petite extension qui contient l'union des extensions des éléments de la classe. On peut exprimer la stabilité d'une classe C dont les éléments sont notés c_i, à l'aide du critère suivant :

$st(C)=card(\left|\cup c_i\right| - \cup\left|c_i\right|)$ Si st(C)=0 donc c est stable.

Effritement d'une classe

C'est le plus petit nombre d'objets symboliques $a \subset C$ dont la réunion des extensions est contenu dans l'extension des éléments d'une classe C tout en s'en écartant le moins possible. Etant données une classe C d'élément c_i et une autre F d'élément f_i, le critère suivant mesure l'effritement :

$E(C)=Min\{card\ F\ /\ \cup\left|f_i\right| = \cup\left|c_i\right|\}$

E(C) est minimum s'il existe un objet F tel que $F=\cup c_i$

4.1.2 Interprétation symbolique

Pour stabiliser l'interprétation des classes de deux partitions, plusieurs méthodes peuvent être utilisées. La méthode d'analyse symbolique pour des données classiques, la méthode

de marquages et généralisations symboliques, et la méthode de comparaison des objets symboliques en utilisant les indices de ressemblances.

Pour étudier la stabilité des interprétations des classes, on utilise l'analyse symbolique traitée sur les données classiques soit dès le départ soit après avoir utilisé une méthode de l'analyse de données classiques pour automatiser l'interprétation. On cherche des objets symboliques complets et d'effritement minimum caractéristiques de chacune des classes, les objets de meilleure stabilité qui minimisent le recouvrement de la partition associée à ces classes.

La méthode MGS Marquages et Généralisations Symboliques [SUM 98,99], présente une autre approche pour étudier la stabilité des interprétations des classes. Cette méthode basant sur l'homogénéité et le critère de discrimination consiste à trouver des descriptions qui généralisent des classes sur l'ensemble des observations classiques. Ces descriptions sont formées par des objets symboliques probabilistes.

En SODAS (Symbolic Official Data Analysis System) la méthode se trouve sous le nom de DSD (Discriminant Symbolic Description). Dans le modèle symbolique, la description d'une classe correspond à une disjonction des objets symboliques.

Une autre façon d'étudier l'homogénéité ou la stabilité de ces interprétations des classes est de comparer leurs descriptions symboliques à l'aide des variables supplémentaires communes aux deux partitions. Si les descriptions sont similaires on peut affirmer leurs stabilités. Pour mesurer la similarité des paires d'objets ou des descriptions symboliques, plusieurs indices ont été développé [GOW 92], [ICH 94], [DEC 98, 00].

SODAS offre la possibilité de choisir entre ces différents indices pour comparer deux objets symboliques de type booléens en utilisant la méthode DI (Distances matrix : Dissimilarities and Matching).

4.6 Cas des données appariées : Même individus, Même variables

Lorsqu'on se trouve dans le cas des données appariées (panels) les partitions provenant d'un même ensemble d'individus se basent sur des questionnaires identiques. Pour tester si la classification de deux partitions est stable, on peut comparer si les moyennes par classes sont identiques. Plusieurs critères et tests proposés dans la littérature [SAP 90]

peuvent être transformées pour ce but de travail. On présente les tests classiques les plus utilisées qui sont les tests d'homogénéité et de Hotelling ou de Mahalanobis.

On procède à une étude basée sur la différence entre les deux tableaux de données X_1 et X_2 appariées $D=X_1-X_2$ en analysant la structure typologique provenant de cette matrice.

4.6.1 Test de Hotelling et distance de Mahalanobis

Une autre façon pour tester si les moyennes par classes sont proches des deux partitions formées des échantillons de n individus à p variables gaussiens indépendants de même matrice de variance Σ, on utilise le test de Hotelling ou la distance de Mahalanobis.

Soient deux tableaux X_1 et X_2 numériques appariés à p dimensions, de même matrice de variance, de n individus et de centres de gravités respectifs g_1 et g_2. Considérons D la différence entre les deux tableaux appariés X_1 et X_2 et de centre de gravité $g= g_1- g_2$. La matrice de variance observée de D est égale à :

$V=1/n\ D'D = 1/n\ (X_1-X_2)'(X_1-X_2)= V_1-V_2+V_{12}-V_{21}$ où V_1, et V_2 sont les matrices de covariances de X_1 et X_2

On veut tester si la moyenne de D est nulle, le test s'écrit :

H0 : μ=0 contre

H1 : $\mu \neq 0$

La distance de Mahalanobis estimée est telle que :

$$D_p^2 = (n-1)(g-\mu)' V^{-1}(g-\mu)$$

Lorsque $\mu=0$, on a : $\dfrac{n-p}{p}(g-\mu)' V^{-1}(g-\mu) = F(p;n-p)$

On présente dans la suite une méthode basant sur la structure de la différence entre les deux tableaux des données appariées.

4.6.2 Classifiabilité de la différence

Notre méthode consiste à procéder l'étude de la matrice de la différence entre les deux tableaux $D=X_1-X_2$, et à effectuer une analyse typologique sur la matrice D. Si aucune

structure de classification sur D n'apparaît, nous pouvons admettre que D est un « bruit » et que les classifications issues de X_1 et X_2 sont semblables.

Le test sera :

H_0 : $D=X_1-X_2=0$ contre

H_1 : $D \neq 0$

Exemple d'application

Prenons deux tableaux provenant des données appariées. On utilise les deux premières procédures de l'algorithme présenté dans le paragraphe (4.3.2) pour créer des données de 1000 individus. Le premier tableau X_1 est formé de 4 variables normales indépendantes, le deuxième X_2 est formé par les mêmes variables de même paramètres, donc on trouve deux tableaux de variables multinormales et de même matrice de variance. On calcul la matrice de différence de ces deux tableaux $D= X_1-X_2$ puis on cherche la structure de classification de D. On utilise la classification hiérarchique par la méthode du critère de Ward.

On utilise la méthode PARTI/DECLA du logiciel SPAD qui fournit la recherche de la meilleure partition en nombre de classes. Le programme propose une partition a deux classes l'une avec la proportion de 99 % la deuxième classe a une proportion de 1 %. Ce qui montre que la manque de structure dans D.

On présente la structure de classification de D pour ce choix de paramètres ainsi que sa distribution graphique.

Statistiques sommaires des variables continues

Libellé de la variable	Effectif	Poids	Moyenne	Ecart-type	Minimum	Maximum
Variable n° 1	1000	1000.00	0.055	0.657	-2.030	5.250
Variable n° 2	1000	1000.00	-0.043	1.314	-8.750	9.270
Variable n° 3	1000	1000.00	0.052	1.295	-9.340	5.070
Variable n° 4	1000	1000.00	0.246	1.785	-9.740	11.290

Tab. 4.13 *Descriptions statistiques des variables de D*

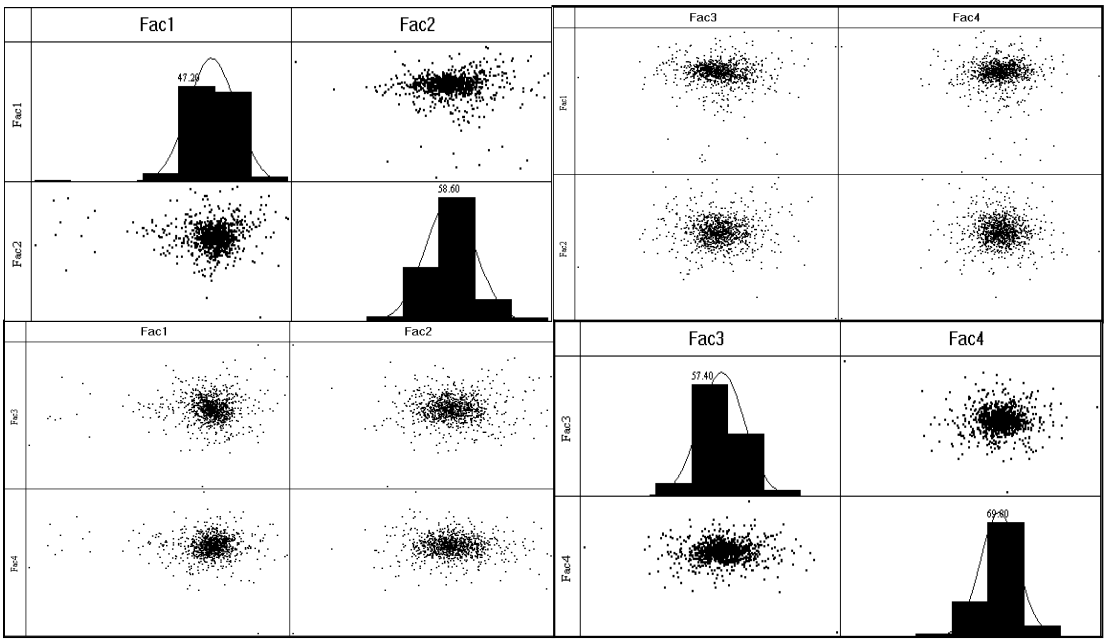


Fig. 4.18 *Nuage de points des individus actives après ACP*

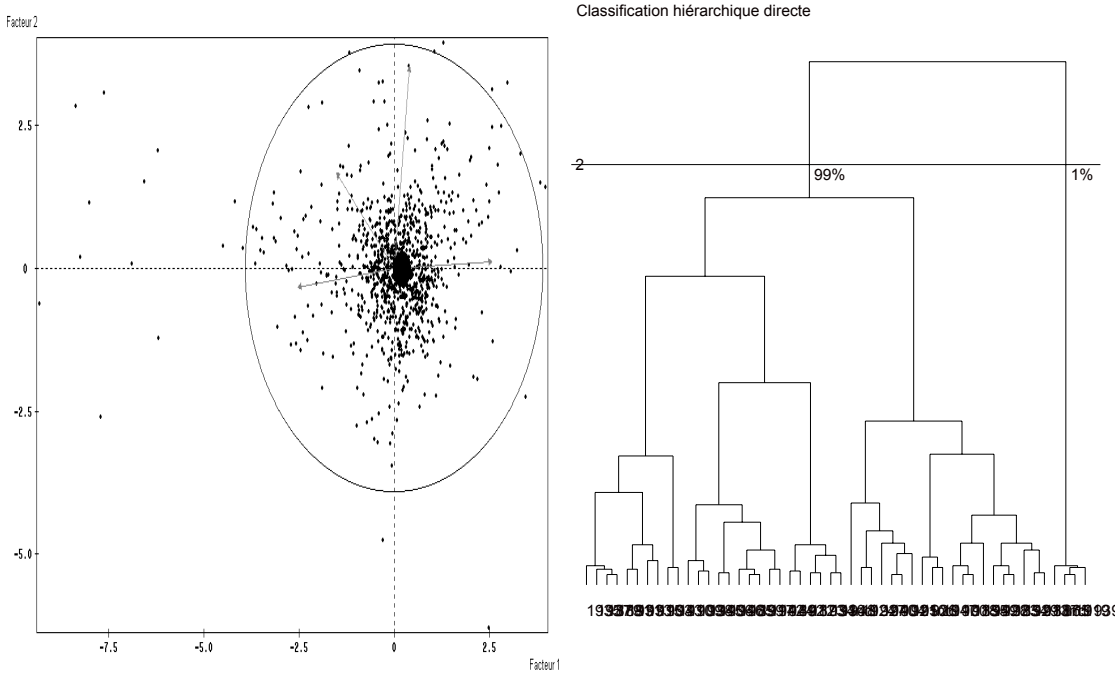


Fig. 4.19 *Représentation des points- classes au barycentre des individus, et le dendrogramme après CAH*

On remarque que la structure hiérarchique de la classification de D est concentrée dans une même classe ce qui confirme donc que le meilleur modèle est celui à une seule classe, d'où la manque de structure de la classification de D.

Par suite, on peut conclure que les classifications issues des deux tableaux sont semblables.

4.7 Conclusion

Nous venons de présenter une étude sur la comparaison des partitions proches ayant des variables différentes pour un même ensemble d'individus. Cette étude sera plus complète si on s'intéresse au cas où on aurait à comparer deux partitions provenant d'un ensemble de données ayant même variables. Ceci fera l'objet du chapitre suivant.

Les tests de stabilité d'une classification ou d'homogénéité sont d'une portée pratiquement réduite comme le souligne [BOC 85], seule une prise en compte précise du modèle probabiliste à une classification permet d'avoir un point de vue critique sur les résultats et permettra de déboucher sur des techniques pertinentes d'évaluation des résultats en classification.

Chapitre 5

Comparaison de partitions de deux groupes d'individus différents décrits par les mêmes variables actives

5.1 Introduction

Le cas d'avoir à comparer des partitions issues de deux échantillons différents se présente fréquemment lors d'enquêtes périodiques d'opinion ou de marché où le même questionnaire est posé à des différents échantillons mais de structure semblable. Il est certes théoriquement possible de tester si les échantillons sont ou ne sont pas significativement différents, mais outre que cela n'est pas facile pour des questionnaires qualitatifs, cela ne répond pas vraiment à la question.

Dans un premier temps, ce chapitre est consacré à la présentation des tests classiques [SAP 90], [LEB 97] de comparaison de deux échantillons. On distingue parmi eux, le test du khi-deux, et le test de Mahalanobis.

La deuxième partie propose une nouvelle méthode de comparaisons par projection des partitions. On applique l'analyse discriminante sur une des deux partitions et on reclasse les individus de l'autre partition. La comparaison sera faite à partir des indices de comparaison des partitions définis au chapitre 2.

Une autre approche pour la comparaison de partitions dans notre cas est définie en utilisant la classification des variables. Il s'agit de trouver les arbres hiérarchiques par les méthodes de classification de variables, et de les comparer à partir des indices de

consensus. Cette méthode a été développée par [ANA 00] et que nous présentons dans la troisième partie de ce chapitre.

La dernière partie traite la stabilité des interprétations des classes des partitions étudiées. On recourra à la comparaison des descriptions symboliques trouvées par application de l'analyse symbolique sur des données classiques ou par utilisation des sorties de la méthode de descriptions des classes (PARTI-DECLA) du logiciel SPAD.

5.2Tests classiques de comparaison de deux échantillons

5.2.1 Proportions des classes : Test du Khi-deux

Considérons le cas où on a à tester si les proportions des classes des partitions n'ont pas varié, sous réserve que les classes sont identiques, au sens où elles ont même signification, d'une partition à l'autre. Pour cela, le test d'homogénéité du khi-deux évoqué précédemment au paragraphe 4.7.1. peut être utilisé.

Soient P_1 et P_2 les partitions de k classes de deux échantillons, à n_1 et n_2 observations formées par les même variables $x_1,\ldots,x_p$. On souhaite tester l'hypothèse suivante:

H_0 : proportions identiques

H_1 :proportions significativement différentes.

On a dans ce cas: $n_1= \sum_{h=1}^{K} n_{1h}$ et $n_2= \sum_{h=1}^{K} n_{2h}$

Pour l'hypothèse H_0, les $p_1, p_2, \ldots,p_k$ représentent les probabilités d'être dans les classes $c_1, c_2,\ldots., c_k$. Il s'agit donc de comparer les effectifs constatés n_{1h} ou n_{2h} aux effectifs espérés n_1p_h et n_2p_h.

On a :

$$d^2=\sum_h \frac{(n_{1h} - n_1p_h)^2}{n_1p_h} + \sum_h \frac{(n_{2h} - n_2p_h)^2}{n_2p_h}$$

On estime les $\hat{p}_h = \frac{n_{.h}}{n_1+n_2} = \frac{n_{1h}+n_{2h}}{n_1+n_2}$, ce qui fait (k-1) estimations indépendantes.

$$d^2 = \sum_h \frac{n_{1h} - \frac{n_1 \times n_{.h}}{n_1 + n_2}}{\frac{n_1 \times n_{.h}}{n_1 + n_2}} + \sum_h \frac{n_{2h} - \frac{n_2 \times n_{.h}}{n_1 + n_2}}{\frac{n_2 \times n_{.h}}{n_1 + n_2}}$$

D^2 est un χ^2 de degré de liberté 2k-2-(k-1) = χ^2_{k-1}, pour deux partitions homogènes.

5.2.2 Comparaison des moyennes des classes :Test de Mahalanobis

Pour la comparaison de deux partitions issues de deux échantillons décrits par les mêmes variables, une autre approche peut-être utilisée, si les variables sont continues. Elle consiste à comparer les moyennes des classes des deux partitions provenant de ces données. Comme nous l'avons vu au paragraphe 4.7.2 le test de Mahalanobis peut-être utilisé pour effectuer cette comparaison.

Pour deux partitions P_1 et P_2 de k classes à n_{1h} et n_{2h} observations pour une classe h on a à tester si Δ_k^2, le carré de la distance de Mahalanobis entre les moyennes des classes des deux partitions, est nul.

La distance estimée de Δ_k^2 est :

$$D_k^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$

Avec W est estimée à partir de deux groupes de matrices de variances V_1 et V_2 des groupes à comparer :

$$W = \frac{n_{1h} V_1 + n_{2h} V_2}{n_{1h} + n_{2h} - 2} = \hat{\Sigma}$$

L'espérance mathématique de D_k^2 est :

$$E(D_k^2) = \frac{n_{1h} + n_{2h} - 2}{n_{1h} + n_{2h} - k - 1} [\Delta_k^2 + \frac{k(n_{1h} + n_{2h})}{n_{1h}.n_{2h}}]$$

Lorsque Δ_k^2=0, donc on a : $\frac{n_{1r} n_{2r} (n_{1r} + n_{2r} - k - 1)}{(n_{1r} + n_{2r}).k.(n_{1r} + n_{2r} - 2)} D_k^2 = F(k; n_{1r} + n_{2r} - k - 1)$

5.3 Projections des partitions

En s'inspirant de l'analyse discriminante, on propose une nouvelle méthode pour la comparaison de deux partitions provenant des échantillons de même variables. Cette méthode consiste à prendre un des deux échantillons comme référentiel et à projeter l'autre échantillon sur les classes de ce référentiel. Nous présentons par la suite les détails de cette méthode ainsi que son application sur des données simulées.

5.3.1 Analyse Discriminante

L'analyse discriminante [HAN 81], [TOM 88], [SAP 90] étudie des données provenant de groupes connus *a priori*. Elle vise deux buts principaux:

- **Description:** Parmi les groupes connus, quelles sont les principales différences que l'on peut déterminer à l'aide des variables mesurées? Pour cet aspect, on cherche à obtenir les projections des moyennes des groupes les plus dispersées possibles et les projections des observations d'un même groupe les plus rapprochées possibles de la projection de la moyenne du groupe. On cherche alors à observer des groupes compacts et distants les uns des autres.
- **Classement:** Peut-on déterminer le groupe d'appartenance d'une nouvelle observation uniquement à partir des variables mesurées? Il s'agit alors de décider dans quelle catégorie il faut l'affecter. Deux approches sont utilisées pour réaliser ce classement : une approche géométrique et une autre probabiliste. Dans l'approche géométrique, il s'agit de calculer la distance entre la nouvelle observation et le centre de chacun des groupes. On classe la nouvelle observation dans le groupe pour lequel cette distance est minimale. Pour l'approche probabiliste, on classe une observation dans le groupe pour lequel la probabilité conditionnelle d'appartenir à ce groupe, étant donnée les valeurs observées, est maximale.

La qualité d'une règle de classement peut se mesurer à l'aide du pourcentage de bien classé sur un échantillon-test ou par validation croisée.

Pour notre cas, nous s'intéressons à l'analyse factorielle discriminante [CEL 94], [LEB 97] qui consiste à rechercher les combinaisons linéaires de p variables qui permettent de

séparer aux mieux les k classes. La première combinaison linéaire sera celle dont la variance interclasse est maximale, afin d'exalter les différences entre les classes, et dont la variance intraclasse est minimale pour que l'étendue dans les classes soit délimitée celle qui discrimine le mieux les classes.

Fonctions linaires discriminantes

Considérons une partition en k classes définies *a priori* par la variable y nominale à k modalités : on calcule les fonctions de classement que l'on utilise ensuite pour reclasser de nouveaux individus. L'approche géométrique d'affectation consiste à choisir la classe dont le centre de gravité qui est le plus proche du point-individu.

$$d^2(e, g_i) = (e - g_i)' W^{-1} (e - g_i) = e' W^{-1} e - 2 g_i' W^{-1} e + g_i' W^{-1} g_i$$

$$\text{Min } d^2(e, g_i) = \text{Max } (2 g_i' W^{-1} e - g_i' W^{-1} g_i)$$

Si on a k groupes donc il faut k fonctions discriminantes. On classe dans le groupe pour lequel la fonction est maximale.

Mais cette approche purement géométrique ne prend pas en compte les probabilités *a priori.* Le modèle bayésien d'affectation permet d'enrichir ce point.

La classe d'affectation de x, l'un des nouveaux individus décrits par les même variables $(x_1,\ldots,x_p)$, sera celle pour laquelle le produit $P(x/I_k).P(I_k)$ est maximal. $P(I_k)$ est la probabilité *a priori* du groupe k, et $P(x/I_k)$ est la probabilité de x sachant que I_k est réalisé. C'est le modèle bayésien d'affectation [LEB 97], lorsque les variables sont normales avec matrice de covariances identique, les fonctions de classement sont linéaires. Notons $f_k(x)$ la densité de probabilité de x connaissant I_k dans le cas multinormal, μ_k et Σ_k désignent respectivement la moyenne et la matrice des covariances théoriques à l'intérieur de groupe I_k. Dans le cas où les distributions de chaque classe ont même matrice de covariances, la densité s'écrit :

$$f_k(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - \mu_k)' \Sigma^{-1} (x - \mu_k)\}$$

l'affectation se fera selon le choix de k qui minimise la fonction de classement $sc_k(x)$ suivant :

$$sc_k(x)= (x-\mu_k)'\Sigma^{-1}(x-\mu_k) \text{-2 Ln } P(I_k)$$

Si de plus les probabilités *a priori* sont égales, la règle de classement coïncide avec la minimisation de la distance de Mahalanobis:

$$sc_k(x)= (x-\mu_k)'\Sigma^{-1}(x-\mu_k)$$

La règle d'affectation bayésienne devient la recherche du centre le plus proche selon cette distance.

Max $p_k f_k(x)$ ⇨ Attribuer x au groupe le plus probable *a posteriori*

Il est alors le maximum de :

$$\text{Max } [\ln p_k - (x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k|$$

- Pour $\Sigma_1= \Sigma_2= \Sigma_3=\ldots= \Sigma$, on attribue x au groupe k tel que :

$$\text{Max}[\ln p_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + x'\Sigma^{-1}\mu_k]$$

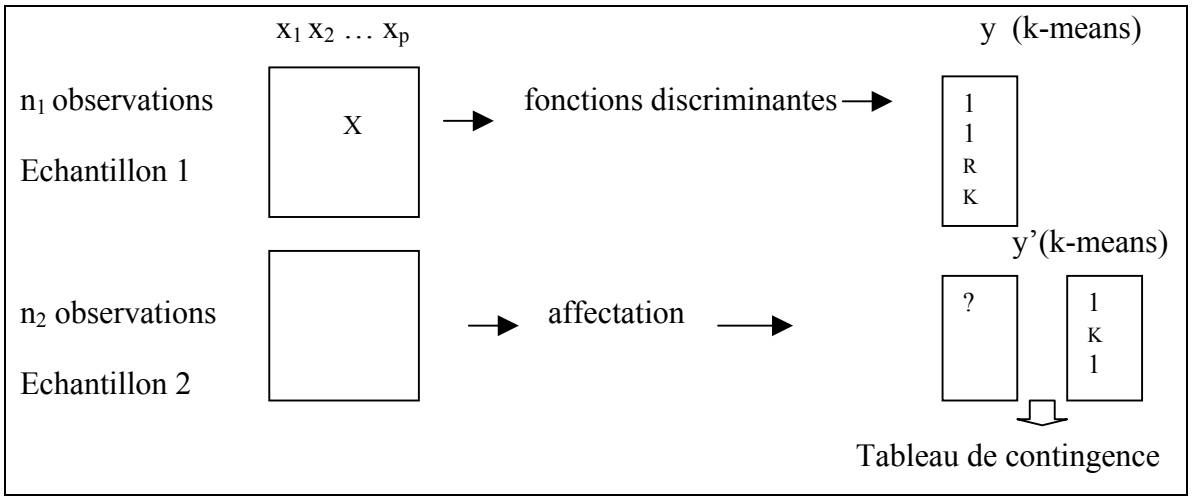
La règle linéaire est équivalente à la règle géométrique si on a l'équiprobabilité après estimation de μ_k par g_k et de Σ par W.

5.3.2 Discrimination sur une partition et reclassement des individus de l'autre partition

Notre méthodologie pour comparer les deux échantillons de mêmes variables est basée sur l'utilisation de la projection des partitions. Pour cela, on réalise la projection de l'un des échantillons sur la partition de l'autre de la façon suivante :

Soit le premier échantillon formé de n_1 observations décrits par p variables $(x_1,\ldots,x_p)$. On cherche la partition de cet échantillon par les méthodes connues de classifications comme les k-means. D'où on trouve la répartition de cet échantillon en k classes définies par la variable y nominale à k modalités.

On définit les fonctions de classement dans le groupe de la première typologie qui sert d'ensemble d'apprentissage.



D'autre part, on cherche une partition pour ce deuxième échantillon par la méthode des k-means. On trouve la répartition en k classes de cet échantillon.

On croise le tableau de contingence formé par les classes nouvelles obtenues par la méthode des k-means, et les classes anciennes reconstituées par l'analyse discriminante. Ce tableau de contingence est alors décrit et analysé par les méthodes de comparaison de deux partitions provenant d'un même ensemble d'individus.

Pour mieux cerner la démarche proposée, on présente dans la suite les détails de cette méthode de projection et son application sur des données simulées.

5.3.3 Algorithme

L'algorithme pour trouver les deux partitions d'un même ensemble de variables se déroule de la manière suivante :

- En utilisant les k-means, on obtient la partition P_1 sur les p variables de la base de données I_1, d'où on en tire les k classes de P_1.
- On joint les deux bases de données de mêmes variables en une seule base de données en ajoutant la numérotation des classes de I_1. On obtient ainsi la base globale I.
- En appliquant l'analyse discriminante linéaire à I, on retrouve les numérotations des classes de la partition P'_2 de I_2 par les fonctions de classement.
- On utilise les k-means pour trouver la partition P_2 de la base de données I_2

- On fabrique le tableau de contingence liant la partition P_2 et celle trouvée par les fonctions de classement P'_2
- Pour étudier la ressemblance de ces deux partitions, on calcule l'un des indices de comparaisons de deux partitions d'un même ensemble d'individus, évoquée précédemment au chapitre 3. En particulier, l'indice Kappa de Cohen et l'indice de Rand.

5.3.4 Simulation

Pour vérifier la pertinence de notre approche nous simulons tout d'abord des partitions proches selon la méthode des classes latentes exposée dans le chapitre 4 (en paragraphe 4.3) de la façon suivante :

- On tire les effectifs des classes latentes selon une loi multinomiale. Pour chaque classe, on tire p variables normales indépendantes. On obtient la première base de données I_1 de N_1 individus.
- De même, on tire les effectifs des classes latentes et pour chaque classe on tire les mêmes p variables normales indépendantes. On trouve la deuxième base de données I_2 de N_2 individus.

On applique cet algorithme en utilisant les logiciels Splus et SAS. On obtient les deux partitions de 500 individus P_1 et P_2 ayant 4 classes chacune à des groupes de 4 variables normales indépendantes. On extrait le tableau de contingence croisant la nouvelle partition P_2 trouvée par la méthode des k-means et la partition P'_2 reconstituée par projection sur la partition P_1. La comparaison de ces échantillons est effectuée selon les deux choix de paramètres présentés au chapitre précédent (Tab. 4.4) et (Tab. 4.5).

Les tableaux de contingence obtenus dans les deux cas de paramètres, représentent les deux partitions provenant d'un même ensemble d'individus : une partition P'_2 qui représente la classification des individus de l'échantillon I_2 après leurs projections sur les classes du premier échantillon I_1 considéré comme référentiel, et une autre partition P_2 qui représente une autre classification des individus de l'échantillon I_2.

- **Premier choix**

Pour le premier choix de paramètres des variables normales indépendantes, le tableau de contingence sera :

P_2 \ P'_2	**1**	**2**	**3**	**4**	P_2 \ P'_2	**2**	**3**	**4**	**1**
1	10	1	0	102	**1**	1	0	102	10
2	121	0	0	0	**2**	0	0	0	121
3	0	123	3	0	**3**	123	3	0	0
4	0	7	132	1	**4**	7	132	1	0

Tab. 5.1 *Tableau croisant P_2 de I_2 par k-means et P'_2 des fonctions discriminantes pour type1 et celui réordonné selon la numérotation du κ maximal*

L'indice kappa, après la permutation pour trouver la numérotation des classes, prend une valeur de 0.941209, et l'indice de Rand a une valeur de 0.957896. Ce qui permet de dire qu'à partir de ces deux valeurs élevées et proches de 1 que les deux partitions sont proches et par suite les deux échantillons présentent des typologies semblables.

- **Deuxième choix**

Le tableau de contingence des paramètres des variables normales obtenus pour le deuxième choix, est donné par :

P_2 \ P'_2	**1**	**2**	**3**	**4**	P_2 \ P'_2	**1**	**3**	**2**	**4**
1	130	0	0	8	**1**	130	0	0	8
2	1	0	140	0	**2**	1	140	0	0
3	0	110	0	3	**3**	0	0	110	3
4	2	8	0	98	**4**	2	0	8	98

Tab. 5.2 *Tableau croisant P_2 de I_2 par k-means et P'_2 des fonctions discriminantes pour Type2 et celui réordonné selon la numérotation du κ maximal*

Dans ce cas, l'indice kappa après permutation prend la valeur de 0.95184, l'indice de Rand est égal à 0.941113. On remarque que ces deux valeurs sont très proches de 1, d'où la similarité entre les deux partitions. Dans ce cas aussi, on peut dire que les deux échantillons des différents individus et de même variables sont stables.

On peut étudier la liaison entre les deux échantillons en utilisant l'indice asymétrique de redondance RI ou τ_b (présenté au chapitre 3). Cet indice mesure la qualité de prédiction de la partition P'_2 sur P_2. Si RI est proche de 1 on a une liaison forte.

La valeur de l'indice RI de la partition P'_2 sur P_2, dans le premier cas, vaut 0.8900734, et dans le deuxième cas, il prend la valeur de 0.8894005. On peut donc conclure, à partir de ces valeurs, que la qualité de prédiction de P'_2 par P_2 est élevée. D'où la forte liaison entre les deux échantillons I_1 et I_2 dans les deux cas de simulations.

Dans le cas où l'échantillon I_2 serait considéré comme référentiel, on projette cette fois ci les individus de I_1 sur les classes de l'échantillon I_2, on trouve le tableau de contingence entre P_1 (trouvée par les k-means) et P'_1 (après projection) dans les deux cas de paramètres.

Pour le premier choix de paramètres, le tableau de contingence est alors:

P'_1 / P_1	**1**	**2**	**3**	**4**
1	2	127	0	0
2	116	1	0	0
3	1	0	127	0
4	0	0	0	126

Tab. 5.3 *Tableau de contingence croisant P_1 de I_1 par k-means et P'_1par projection de I_1 sur les classes de I_2 supposé comme référentiel.*

Pour ce cas, l'indice de redondance RI est égal à 0.9789137. d'où la forte liaison entre les deux partitions. On peut donc conclure que les deux échantillons sont proches.

Pour le deuxième cas, on obtient :

P_1 \ P'_1	**1**	**2**	**3**	**4**
1	2	0	11	107
2	0	123	0	0
3	119	0	0	0
4	0	0	136	2

Tab. 5.4 *Tableau de contingence croisant P_1 de I_1 et P'_1pour le 2^{ieme} type de paramètres.*

L'indice de redondance RI prend une valeur de 0.9263896 très proche de 1 ce qui montre encore une fois la ressemblance des deux partitions.

Il est à noter que les résultats de la projection de l'échantillon I_2 sur les classes de l'échantillon référentiel I_1 et de I_1 sur les classes de I_2 ne sont pas identiques dans les deux types de choix de paramètres. Ceci impose la nécessité de l'utilisation d'un indice asymétrique tel que l'indice de redondance RI pour la comparaison de deux partitions provenant d'un ensemble de même variables et de différents individus. D'autant plus que cet indice ne nécessite pas de permuter les classes avant son utilisation comme le cas de l'indice kappa de Cohen.

5.4 Autre approche par la classification des variables

Lorsqu'on a différents individus mais des variables identiques, on peut inverser la problématique en comparant les deux classifications de variables que l'on peut obtenir. Lorsque les classifications des variables sont semblables, cela implique de manière duale que la structuration des individus est comparable (classes de même signification, mais de poids éventuellement différent). On utilise alors une des classifications hiérarchiques.

On est conduit à un problème de comparaison de classifications hiérarchiques. Il faut se donner des modèles probabilistes adaptés donc une distribution de hiérarchies sous l'hypothèse de tirages aléatoires dans une population d'individus. De tels travaux

existent, basés sur des mesures de consensus, mais semblent encore difficiles à appliquer [SOK 88].

Dans cette partie, on présente une méthode pour comparer deux classifications hiérarchiques des variables.

5.4.1 Méthodes de Classification de variables

Les techniques de classifications hiérarchiques de variables

Deux types de techniques sont utilisés pour la classification de variables [NAK 00], les techniques basées sur l'algorithme agglomératif et les techniques basées sur l'algorithme divisif.

Les techniques d'algorithme agglomératif relèvent de la classification ascendante hiérarchique. Elles requièrent l'utilisation :

- D'un indice de similarité entre variables dépendant de la nature des variables,
- D'un critère d'agrégation qui permet d'ériger un système de classes de variables emboîtées.

La procédure CLUSTER dans le logiciel SAS réalise cette technique où 11 critères d'agrégations peuvent être choisis.

Les techniques basées sur un algorithme divisif (présenté au chapitre 2) sont fondées sur l'utilisation d'un critère de division d'un sous-ensemble de variables [CHA 97]. La procédure VARCLUS du logiciel SAS [SAS 94] fournit une telle méthode de classification. Cette procédure fournit une classification basée sur la matrice des corrélations (cas des variables de même poids) ou sur la matrice des variances-covariances (cas où les variables doivent avoir plus d'importance quand leurs variances sont grandes). La partition obtenue est telle que les variables d'une même classe sont aussi corrélées entre elles que possible et deux variables quelconques de deux classes différentes sont les moins corrélées possible. Les classes de variables sont ainsi construites de manière à rendre maximum la variance expliquée par la première composante principale de chaque classe de la partition.

Récemment, une autre méthode de classification de variables basée sur les composantes latentes a été proposée par Vigneau E. [VIG 03]. La classification de variables autour des composantes latentes est considérée comme étant un moyen d'organiser des données multivariées dans des structures significatives. Cette méthode distingue deux cas selon que le signe de la corrélation est important ou non (soit on utilise r soit r^2). La stratégie consiste à faire une classification hiérarchique puis à effectuer une méthode de partitionnement. Les deux algorithmes cherchent à maximiser le même critère qui offre la possibilité de savoir quelles variables dans chaque classe sont reliées à une variable latente associée à la classe. On verra dans la suite un exemple d'application de cette méthode sur des données simulées.

Indice de Similarité entre Variables

Pour effectuer la classification hiérarchique des variables, on a besoin d'une matrice de similarités trouvée à partir d'une mesure de similarités entre variables. La notion de cette mesure est due à la nature des variables. Ce paragraphe présente quelques indices de similarités utilisés pour les différentes natures des variables.

- **Données numériques**

La similarité entre deux variables j et j' d'un tableau de terme général x_{ij} est fournie par le coefficient de corrélation linéaire de « Bravais- Pearson » bien connu:

$$r_{jj'} = \frac{\sum_i (x_{ij} - \overline{x_j})(x_{ij'} - \overline{x_{j'}})}{\left[\sum_i (x_{ij} - \overline{x_j})^2 \sum_i (x_{ij'} - \overline{x_{j'}})^2\right]^{1/2}}$$

Avec n est le nombre d'observations et $\overline{x_{j'}}$ est la moyenne de la variable j'.

- **Données de fréquences**

A partir du tableau de fréquences de terme général f_{ij}, on définit la distance entre deux colonnes j et j' comme la distance de khi-2 associé à f_i. L'expression de cette distance associé à f_i est obtenue à partir de la distance du khi-deux entre deux individus en changeant i en j, soit :

$$d^2_{jj'} = \sum_i \frac{1}{f_i} \left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}}\right)^2$$

- **Données ordinales**

Pour étudier la liaison entre deux variables ordinales, Spearman a proposé de calculer le coefficient de corrélation sur les rangs afin de comparer les classements issus de ces deux variables. Pour deux structures hiérarchiques, on peut utiliser ce coefficient pour comparer l'ordre d'agrégation trouvé après la CAH des variables des deux données. Le coefficient de Spearman est défini par :

$$r_s = 1 - \frac{6\sum_j d_j^2}{p(p^2-1)}$$

où p= le nombre de variables dans une structure hiérarchique,

d_j=la différence des rangs d'une même variable selon les deux classements

Cette définition nous indique que :

Si r_s =1 alors les deux classements sont identiques

r_s =0 alors les deux classements sont indépendants

r_s =-1 alors les deux classements sont inverses l'un de l'autre

Pour savoir si la valeur trouvée de r_s est significative, on se reportera à la table du coefficient de Spearman [SAP 90]. La région critique sera $|r_s| \succ r'_s$:

- **Données binaires**

L'indice de similarité entre deux variables binaires est calculé à partir du tableau de contingence (2x2) obtenue en croisant les deux variables. L'indice de similarité le plus courant est le Φ^2 de Pearson qui prend des valeurs comprises entre 0 et 1 et qui est obtenu á partir du khi-deux de contingence.

$$\Phi^2_{jj'} = \frac{\chi^2_{jj'}}{n} \text{ où } \chi^2_{jj'} = \frac{n(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

- **Données nominales**

Pour deux variables nominales, l'indice de similarité est calculé à partir du tableau de contingence croisant les deux variables dont le nombres de modalités sont respectivement

p et q. L'indice de similarité est le coefficient $C_{jj'}$ de Cramer [KEN 61], obtenu à partir du $\Phi^2_{jj'}$.

$$C_{jj'} = \frac{\Phi^2_{jj'}}{\min(p-1, q-1)}$$

- **Données mixtes**

Dans ce cas de variables mixtes, on peut transformer le tableau de départ en tableau disjonctif complet, en discrétisant les variables numériques, à partir duquel on calcul les distances du khi-deux entre deux colonnes.

- **Coefficient d'affinité**

Le coefficient d'affinité a été introduit en statistique inférentielle par Matusita K. [MAT 55] et développé par Bacelar- Nicolau H. [BAC 85, 02], et utilisé comme coefficient de ressemblance. Il mesure la tendance monotone entre les racines carrées des vecteurs profils ou probabilités.

Le coefficient d'affinité simple entre les paires des variables x_j et $x_{j'}$ pour j et j' $=\{1,\ldots,p\}$, peut être définie, dans le cas d'équiprobabilité, de la façon suivante :

$$C_\alpha = \sum_i \sqrt{\frac{x_{ij}.x_{ij'}}{x_{.j}x_{.j'}}}$$

où $$x_{.j} = \sum_{i=1}^{n} x_{ij} \text{ et } x_{.j'} = \sum_{i=1}^{n} x_{ij'}$$

le coefficient d'affinité généralisé [BAC 03] pour les variables réelles qui peuvent être négatives est défini par :

$$C_\alpha = \sum_i \sqrt{\left|\frac{x_{ij}.x_{ij'}}{x_{.j}x_{.j'}}\right|}$$

avec $x_{.j} = \sum_{i=1}^{n} \left|x_{ij}\right|$ et $x_{.j'} = \sum_{i=1}^{n} \left|x_{ij'}\right|$ ($|\ |$ désigne la valeur absolue)

Il est symétrique et prend ses valeurs dans l'intervalle [0,1] ; il est égal à 1 si les deux vecteurs sont identiques ou proportionnels et nul s'ils sont orthogonaux. On remarque que ce coefficient est indépendant de la taille des données et des variables.

Le coefficient d'affinité peut être appliqué sur tout type de données : il génère le coefficient d'Occhiai pour les données binaires, et peut être adapté sur les données de fréquences, des données réelles quantitatives, dans ce cas on utilise le coefficient d'affinité généralisé [BAC 03], de distributions normale et uniforme. [BAR 00] a définit le coefficient d'affinité pour mesurer la similarité ente histogrammes, entre variables symboliques modales et de type intervalle.

Plusieurs travaux basés sur ce coefficient et ses extensions appliquent des techniques et méthodes d'analyse multivariée comme l'approche probabiliste dans l'analyse de classification non hiérarchique symbolique [BAC 02], l'effet des données manquantes dans la classification des variables hiérarchiques [SIL 02], ou le problème de validation [SOU 02]. Il est introduit dans plusieurs logiciels tel que SODAS [SOU 02].

Il existe une version probabiliste de ce coefficient pour valider les résultats d'une classification dans un modèle probabiliste, le coefficient VAL (validity linkage) qui peut être trouvé dans [LER81], [BAR88], [NIC 88].

Critère d'agrégation pour la classification des variables

Une fois la mesure de similarité d est choisie, il reste à choisir un critère d'agrégation pour ériger l'arbre de la CAH (Classification Ascendante Hiérarchique). Ce critère est une règle de calcul des écarts entre deux sous–ensembles h et h' disjoints de l'ensembles des éléments à classer. Parmi les nombreux critères d'agrégation, les critères usuels, sont les suivants :

- Le critère du saut minimal (Single Linkage), dû à Jardine et Sibson [JAR 71] crée, en agrégeant en priorité les paires de classes entre lesquelles l'écart D_{min} est le plus petit, des classes bien séparées entre elles. Il est définit par :

$$D_{min}(h,h') = \inf\{d(i,i');i \in h, i' \in h'\}$$

Ce critère présente un intérêt en reconnaissance des formes puisqu'il permet de détecter des classes longues et sinueuses.

- Critère du diamètre (complete linkage), c'est la plus grande distance entre un point de h et un point de h'. Ce critère dû à Sorensen [SOR 48] crée, en agrégeant en priorité les paires des classes entre lesquelles l'écart est le plus faible, des classes compactes. Noté D_{diam}, il est définit par :

$$D_{diam}(h,h') = \sup\{d(i,i');i \in h,i' \in h'\}$$

Contrairement à D_{min}, ce critère ne possède pas de bonnes propriétés et n'est pas utilisé en pratique [NAK 00].

- Critère de la moyenne (average linkage), c'est la moyenne des distances entre un point x_i de masse m_i et un point $x_{i'}$ de masse $m_{i'}$, chaque distance $d(x_i,x_{i'})$ ayant pour masse le produit $m_i m_{i'}$, les masses des classes h et h' sont respectivement m_h et $m_{h'}$. Il est dû à Sokal et Michener [SOK 58], et peut être considéré comme un compromis entre D_{min} et D_{diam}. Il est représenté sous la façon suivante :

$$D_{moy}(h,h') = \frac{1}{m_h m_{h'}} \sum_{i \in h, i' \in h'} \{m_i m_{i'} d(i,i')\}$$

Il peut être utilisé pour classer des individus d'un ensemble dont les masses sont différentes au départ.

5.4.2 Comparaison de classifications hiérarchiques

On présente une méthode de comparaison de deux classifications hiérarchiques des données de même variables. Cette méthode est basée sur la combinaison des structures hiérarchiques, considérant l'ordre de l'agrégation et non pas les niveaux. On peut la considérer comme une méthode de consensus.

[SIL 02] a proposé autre que celle-ci une méthode basée sur la moyenne des matrices de similarités, pour étudier des classifications de données manquantes utilisant les matrices obtenues par imputation multiple.

Algorithme

La procédure de la méthode de comparaison de classification hiérarchique des variables est la suivante :

1. Pour chaque matrice de données $I_1(n, p)$ et $I_2(n, p)$, on détermine sa matrice de similarité,
2. Sur chaque matrice de similarité, on utilise la méthode d'agrégation choisie, pour obtenir une structure hiérarchique, représenté par un dendogramme,
3. A chaque structure hiérarchique, est associé une matrice ultramétrique qu'on détermine,
4. On calcul le coefficient de Spearman r_s entre les deux ultramétriques. Si r_s est égal à 1, les structures sont identiques, on se reporte à la table du coefficient de Spearman où la valeur critique est r'_s :

   $|r_s| \succ r'_s$: les deux structures de classifications hiérarchiques ne sont pas les mêmes, mais les deux ultramétriques sont corrélées significativement.

   $|r_s| \prec r'_s$: les deux structures des classifications hiérarchiques sont significativement différentes.

[SIL 02] a montré, par simulation, que les meilleurs résultats sont obtenus avec la méthode associée au coefficient d'affinité avec les critères d'agrégations D_{moy} et D_{min} et que le coefficient d'affinité C_α est plus robuste que le coefficient de corrélation linéaire r.

Simulation

Pour appliquer l'algorithme par simulation, on cherche des données de même ensembles de variables basées sur des structures probabilistes. On crée deux tableaux de données simulées de 500 individus issus de 4 distributions multinormales selon les deux choix de paramètres présentés dans le chapitre précédent (Tab. 4.4 et Tab. 4.5).

On utilise le coefficient d'affinité C_α pour trouver les matrices de similarités des deux tableaux. Les structures hiérarchiques des deux tableaux I_1 et I_2 sont réalisées à partir du critère d'agrégation de la moyenne D_{moy}.

- **Premier choix**

Pour ce choix de paramètres, la matrice de similarités des variables et les structures hiérarchiques des deux tableaux de données I_1 et I_2, sont :

$$\begin{pmatrix} 1 & . & . & . \\ 0.8464 & 1 & . & . \\ 0.8808 & 0.8727 & 1 & . \\ 0.7909 & 0.8003 & 0.8633 & 1 \end{pmatrix} \text{ et } \begin{pmatrix} 1 & . & . & . \\ 0.857 & 1 & . & . \\ 0.865 & 0.899 & 1 & . \\ 0.777 & 0.846 & 0.856 & 1 \end{pmatrix}$$

Tab. 5.5 *Les matrices de similarités S_1 et S_2 de I_1 et I_2 respectivement*

Les dendrogrammes, trouvées par la méthode de classification hiérarchique des variables et en utilisant le critère d'agrégation de la moyenne, des deux matrices de similarités S_1 et S_2, ont la forme suivante :

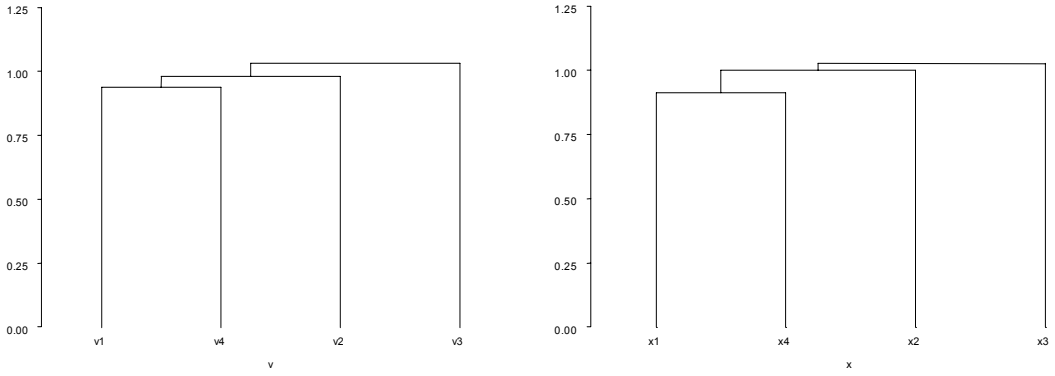


Fig.5.1 *Dendrogramme des deux matrices de similarités S_1 et S_2*

On obtient les matrices suivantes des distances ultramétriques des deux dendrogrammes:

	v_1	v_2	v_3	v_4
v_1	0			
v_2	0.9812	0		
v_3	1.0317	1.0317	0	
v_4	0.9383	0.9812	1.0317	0

	x_1	x_2	x_3	x_4
x_1	0			
x_2	1.000856	0		
x_3	1.026728	1.026728	0	
x_4	0.913269	1.000856	1.026728	0

Tab. 5.6 *Les ultramétriques de S_1 et S_2 respectivement*

Pour comparer les résultats des deux classifications de variables, on utilise le coefficient de Spearman r_s entre les deux ultramétriques. On trouve une valeur de coefficient de Spearman égal à 1 d'où la structure identique des deux classifications, comme le montre la figure (Fig. 5.1)

Deuxième choix

En utilisant la même procédure mais pour ce deuxième choix de paramètres des variables multinormales, on trouve les matrices de similarités des données I_1 et I_2 suivantes :

$$\begin{pmatrix} 1 & . & . & . \\ 0.84627 & 1 & . & . \\ 0.82425 & 0.87981 & 1 & . \\ 0.85809 & 0.89005 & 0.89105 & 1 \end{pmatrix} \text{ et } \begin{pmatrix} 1 & . & . & . \\ 0.851 & 1 & . & . \\ 0.849 & 0.888 & 1 & . \\ 0.874 & 0.899 & 0.905 & 1 \end{pmatrix}$$

Tab. 5.7 *Les matrices de similarités S_1 et S_2 de I_1 et I_2 pour le deuxième choix*

L'application du critère de la moyenne, donne les structures hiérarchiques des matrices de similarités suivantes:

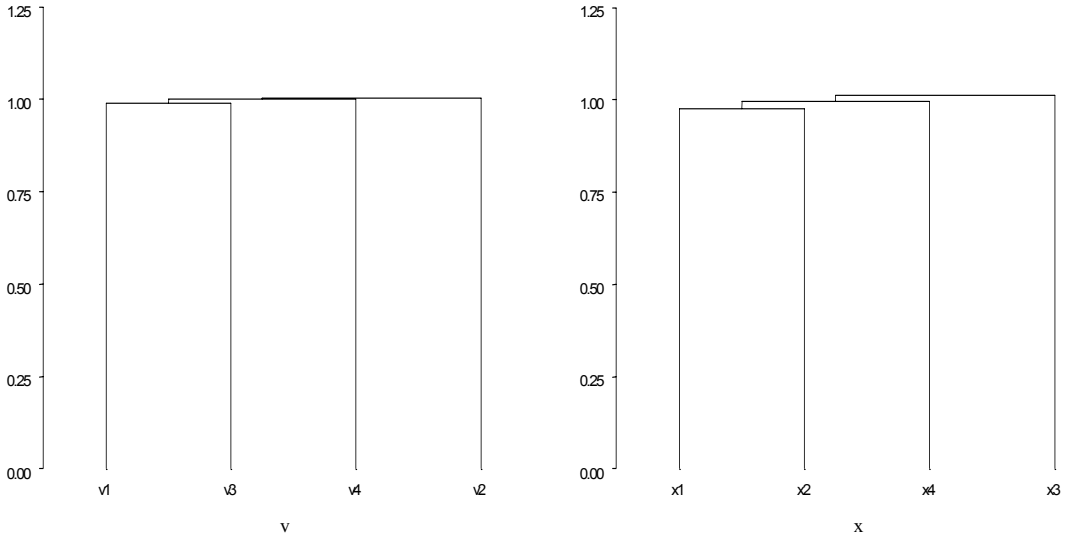


Fig.5.2 *Dendrogramme des deux matrices de similarités S_1 et S_2 du deuxième choix*

Les matrices des distances ultramétriques dans ce cas sont les suivantes :

	v_1	v_2	v_3	v_4
v_1	0			
v_2	1.0028	0		
v_3	0.9907	1.0028	0	
v_4	1.0005	1.0028	1.0005	0

	x_1	x_2	x_3	x_4
x_1	0			
x_2	0.9763	0		
x_3	1.0116	1.0116	0	
x_4	0.9941	0.9941	1.0116	0

Tab. 5.8 *Les ultramétriques de S_1 et S_2 au deuxième choix*

De même, pour ce choix de paramètres, le coefficient de Spearman calculé à partir de l'ordre d'agrégation des structures hiérarchiques, est égale à 1. On peut donc conclure à l'identité des classifications de ces deux données simulées (Fig. 5.2)

5.4.3 Comparaison à partir de VARHCA de Vigneau

Afin de comparaison deux classifications hiérarchiques issues des données de même variables, on considère la méthode de classification de variables proposée par Vigneau et Quannari [VIG 03]. On applique le macro VARHCA écrit par Vigneau dans un exemple d'application.

Exemple d'application

On crée deux tableaux de données simulées de 500 individus issus de 4 distributions multinormales selon le premier choix de paramètres présentés dans le chapitre précédent (Tab. 4.4). On applique le macro VARHCA sur les deux données pour trouver les deux structures hiérarchiques correspondant aux deux tableaux de données. Pour comparer ces structures, on calcul le coefficient de corrélation des deux ultramétriques trouvées. Les dendrogrammes des deux tableaux sont représentées dans la figure suivante :

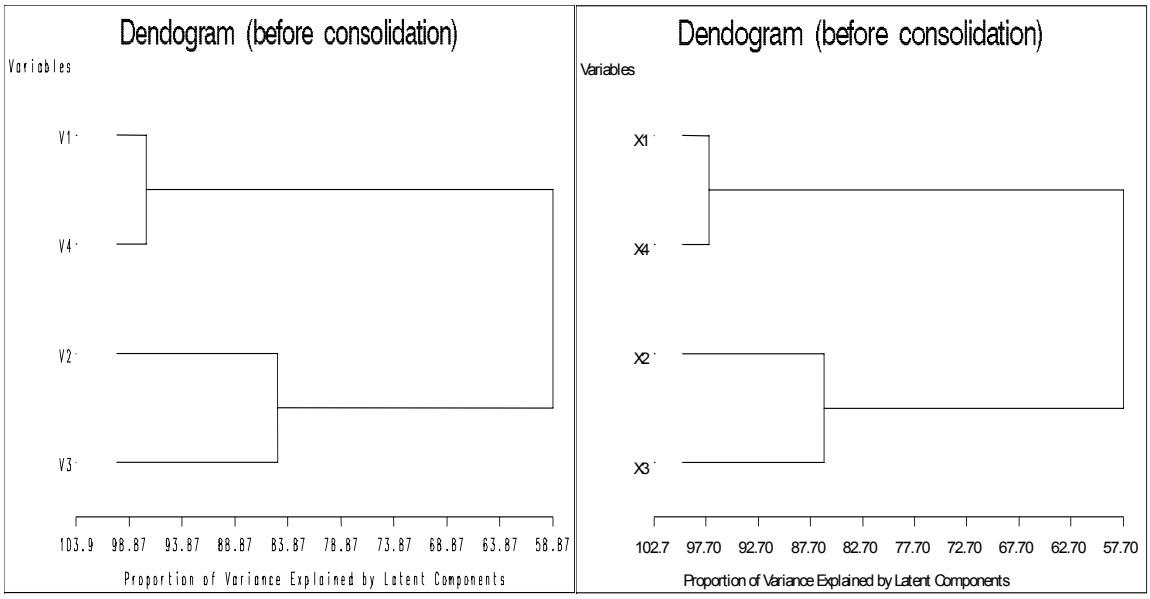


Fig. 5.3 *Dendrogrammes de I_1 et I_2*

Les matrices des distances ultramétriques dans ce cas sont les suivantes :

	v_1	v_2	v_3	v_4
v_1	0			
v_2	71.98	0		
v_3	71.98	34.4	0	
v_4	7.5086	71.98	71.98	0

	x_1	x_2	x_3	x_4
x_1	0			
x_2	81.08	0		
x_3	81.08	31.36	0	
x_4	7.23	81.08	81.08	0

Tab. 5.9 *Les ultramétriques de I_1 et I_2*

La valeur du coefficient de Spearman calculé à partir de l'ordre d'agrégation de ces structures hiérarchiques, est égale à 1. Donc encore on peut déduire l'identité des classifications de ces deux données simulées.

5.5 Stabilité des interprétations

Il est important de vérifier si la signification des classes est restée la même, pour étudier la stabilité des interprétations des classes trouvées par une classification des données. On compare alors les descriptions statistiques des classes ou les descriptions symboliques des classes.

5.5.1 Comparaison des descriptions statistiques

Une des méthodes d'étude de la stabilité des interprétations est basée sur la comparaison des descriptions statistiques des classes. Ces descriptions sont fondées généralement sur des comparaisons de moyennes ou de pourcentage à l'intérieur des classes avec les moyennes ou les pourcentages obtenus sur l'ensemble des éléments à classer.

On peut caractériser soit chaque classe d'une partition, soit globalement la partition elle-même. Tous les éléments disponibles (actifs et illustratifs) peuvent intervenir dans la caractérisation: les modalités des variables nominales, les variables nominales elles-mêmes, les variables continues, les fréquences et les axes factoriels.

Les éléments caractéristiques sont classés par ordre d'importance à l'aide d'un critère statistique ("valeur-test"), auquel est associée une probabilité [MOR 84] permettant d'opérer un tri sur les variables et de designer les variables les plus caractéristiques: plus la valeur-test est grande, plus la probabilité est faible, plus l'élément est caractéristique. La valeur-test mesure l'écart entre les valeurs relatives à la classe et les valeurs globales, elle constitue de simples mesures de similarité entre les variables et les classes.

Pour les variables continues, on compare $\overline{X}_h$ la moyenne d'une variable X de la classe h, à la moyenne générale $\overline{X}$ et on évalue l'écart en tenant compte de la variance $s_h^2(X)$ de cette variable dans la classe. La valeur- test $t_h(X)$ est alors :

$$t_h(X)=\frac{\overline{X}_h-\overline{X}}{s_h(X)}$$

avec

$$s_h^2(X)=\frac{n-n_h}{n-1}\frac{s^2(X)}{n_h}$$

$t_h(X)$ suit approximativement une loi normale centrée réduite. L'interprétation se fait sur variables communes supplémentaires aux deux données. Les variables sont d'autant plus intéressantes que les valeurs-test associées sont fortes en valeur absolue. On peut alors classer selon leur niveau de significatif. On peut tester alors si, pour les deux données de mêmes variables, les variables significatives d'une classe ont les mêmes répartitions[1].

[1] Le logiciel SPAD contient la procédure DECLA qui permet de décrire les partitions obtenues par la procédure PARTI

5.5.2 Comparaison des descriptions symboliques

Une autre méthode pour mesurer la stabilité des interprétations revient à comparer les descriptions symboliques (chapitre 2) des partitions provenant de deux données de même variables. La comparaison est faite en utilisant les indices de similarités pour les variables supplémentaires communes. Les interprétations des partitions sont stables lorsqu'on a une forte similarité entre les descriptions symboliques des classes.

On cherche les deux partitions des deux données de même variables par la méthode divisive de classification [CHA 98]. Cette méthode permet de définir les descriptions des classes trouvées et dont le résultat est un dendrogramme ou arbre de décision. [MEH 03] a proposé une méthode pour trouver des descriptions des classes des partitions en optimisant simultanément le critère de discrimination et le critère d'homogénéité. Chaque classe est décrite par une conjonction des propriétés caractéristiques.

Pour comparer les descriptions symboliques des classes de deux partitions, on utilise les fonctions de comparaisons proposées par De Carvalho [DEC 98] de la façon suivante :

Supposons a=[$Y_1 \in A_1$]^[$Y_2 \in A_2$]^[$Y_2 \in A_2$]]^[$Y_3 \in A_3$].....^[$Y_p \in A_p$] et b=[$Y_1 \in B_1$]^[$Y_2 \in A_2$]^[$Y_2 \in A_2$]]^[$Y_3 \in A_3$].....^[$Y_p \in A_p$] deux objets symboliques booléens, le calcul de la fonction de comparaison pour chaque variable Y_j se base sur le tableau d'agrément suivant :

	Accord	Désaccord	totale
Accord	$M(A_j \cap B_j)=a$	$M(A_j \cap c(B_j))=d$	$M(A_j)$
Désaccord	$M(c(A_j) \cap B_j)=c$	$M(c(A_j) \cap c(B_j))=b$	$M(c(A_j))$
totale	$M(B_j)$	$M(A_j)$	

Tab. 5.10 *Tableau d'accord entre mesure*

Pour un sous-ensemble V_j on a :

$$M(V_j)= \begin{cases} \left|\overline{v}_j - \underline{v}_j\right| \text{ si } Y_j \text{ est conitnue et } V_j = [\overline{v}_j, \overline{v}_j] \\ \left|V_j\right| \text{ si } Y_j \text{ est entier, nominale ou ordinal} \end{cases}$$

De Carvalho [ESP 00] a proposé les fonctions de comparaisons comme extension des mesures de similarité définies pour les variables binaires classiques comme celle de Jaccard,..etc.

Si la fonction de comparaison est proche de 1, on peut conclure que les interprétations de la classification sont stables.

5.5.3 Identification des classes

Pour identifier les classes des partitions, on peut estimer deux modèles de classes latentes. Ces modèles sont des modèles particuliers de mélanges de distributions [EVE 81]. En utilisant l'algorithme EM (Réf. Chapitre 1), les paramètres sont estimés par le maximum de vraisemblances. Pour les deux échantillons, après avoir déterminé le nombre de classes k basé sur la plus petite valeur des critères AIC ou BIC, on estime par l'algorithme EM, π_h les probabilités *a priori* d'appartenir à la classe latente h. Lorsque les variables observées sont binaires on calcule les p_{jh} la probabilité d'avoir une réponse oui pour un individu i de la classe latente h à chacune des p variables conditionnelles aux classes latentes. Il suffit alors de comparer les vecteurs de p_{jh} dans les deux partitions pour tester l'identité des deux modèles.

5.6 Conclusion

Dans ce chapitre, nous venons de présenter les méthodes et les tests classiques de comparaison des partitions provenant des données de mêmes variables mais de différents individus. Une nouvelle méthode de comparaisons est proposée basée sur la projection des partitions. Une autre approche pour aborder la comparaison est proposée par utilisation de la classification des variables. Enfin, la stabilité des interprétations des classes des partitions a été présentée.

Le chapitre suivant sera consacré à l'application sur des données réelles de ces différentes approches évoquées jusqu'à présent.

Chapitre 6

Applications

6.1Introduction

Dans le but de valider l'étude présentée dans les deux derniers chapitres, ce chapitre est consacré à l'application des différents algorithmes sur des données réelles.

Au chapitre 4, on cherche à comparer deux partitions provenant d'un même ensemble d'individus décrits par deux ensembles de variables pour tester si elles sont proches ou non en utilisant la méthodologie de profils latents. On s'intéresse à tester la stabilité des classes et de leurs interprétations pour les deux partitions.

L'application du chapitre 5 s'effectue selon les deux démarches suivantes : la première consiste à comparer les partitions de deux groupes d'individus décrits par les mêmes variables en appliquant la méthodologie de comparaisons par projection de partitions. La deuxième approche consiste à comparer les deux partitions en utilisant la classification de variables.

Notre démarche consiste donc à traiter les mêmes données provenant d'une enquête selon les deux procédures suivantes :

- Même ensemble d'individus avec deux groupes de variables: elle consiste à diviser les données de l'enquête horizontalement afin d'obtenir deux groupes de variables pour le même ensemble d'individus.
- Deux échantillons d'individus ayant même ensemble de variables : elle consiste à diviser les données de l'enquête verticalement pour obtenir deux échantillons d'individus ayant mêmes variables.

Plusieurs logiciels sont utilisés lors du traitement des données :

- Le logiciel Splus est utilisé pour appliquer la méthode des k-means, permuter les données de base, et réaliser les algorithmes qui ont été proposés.
- Le logiciel SPAD est utilisé pour la méthode de la classification hiérarchique ascendante, la méthode PARTI-DECLA pour choisir le nombre de classes des partitions et pour la description des classes.
- Le logiciel SAS est utilisé pour l'analyse discriminante linéaire et pour la classification de variables.
- Le logiciel LatentGOLD pour estimer le modèle de profils latents.

6.2 Description des données

On utilise les données de l'enquête effectuée sur les conditions de vie et aspirations des Français [LEB 87]. Les données comprennent 1000 individus et 52 variables. On s'intéresse aux variables d'opinions traitant la qualité de vie des français, le mariage, la famille, et les enfants, soit 14 variables actives et 6 variables illustratives. Après suppression des données manquantes il reste 624 individus.

Pour la première procédure, on a deux groupes de variables : l'un traitant l'opinion des français sur le mariage, les familles et les enfants, et l'autre traitant l'opinion sur la qualité de vie. Du fait que ces variables ont des modalités nombreuses et ordonnées, on les considère comme étant continues.

Pour le premier groupe, on prend les variables actives suivantes :

```
VARIABLES CONTINUES ACTIVES
     7 VARIABLES
-------------------------------------------------------------------------------------------
   1 . opinion a propos du mariage                                    ( CONTINUE )
   2 . comparée votre santé  a votre age                              ( CONTINUE )
   3 .  la crèche est une mode de garde                               ( CONTINUE )
   4 . la mère au foyer est une mode de garde                         ( CONTINUE )
   5 . nombre d'enfants idéales                                       ( CONTINUE )
   6 . revenu personnel souhaite                                      ( CONTINUE )
   7 . revenu minimum d'une famille                                   ( CONTINUE )
-------------------------------------------------------------------------------------------
```

Pour le deuxième groupe, les variables actives sont :

```
 VARIABLES CONTINUES ACTIVES
     7 VARIABLES
-------------------------------------------------------------------------------------------
   1 . opinion sur le cadre de vie                                    ( CONTINUE )
   2 . opinion sur le fonctionnement de la justice                    ( CONTINUE )
   3 . estimation du salaire d'ingénieur                              ( CONTINUE )
   4 . estimation du salaire d'un médecin                             ( CONTINUE )
   5 . évolution du niveau de vie                                     ( CONTINUE )
   6 . heure de coucher                                               ( CONTINUE )
   7 . nombre de jour de vacance                                      ( CONTINUE )
-------------------------------------------------------------------------------------------
```

Pour les deux groupes, les variables illustratives communes sont :

```
VARIABLES ILLUSTRATIVES
   2 VARIABLES CONTINUES ET 4  VARIABLES QUALITATIVES
-------------------------------------------------------------------------------------------
  12 . age                                                            ( CONTINUE )
  13 . age de fin d'étude                                             ( CONTINUE )
   8 . la famille est l'endroit ou on sent bien                       (   3 MODALITES )
   9 . la préservation de l'environnement est une chose               (   5 MODALITES )
  10 . sexe                                                           (   2 MODALITES )
  11 . profession                                                     (  19 MODALITES )
```

Pour la deuxième procédure, on divise les données en deux parties égales par tirage au sort équiprobable sans remise. On obtient ainsi deux échantillons d'individus ayant même variables. Il est à noter que pour trouver une distribution de l'indice de Rand à partir de l'application de la projection de partitions, on répète 50 fois cette démarche.

6.3 Comparaison des partitions ayant même individus

On effectue une classification hiérarchique par le critère de Ward pour connaître le nombre de classes indispensable pour appliquer la méthode des k-means. Pour cela, on peut choisir selon la coupure sur les dendrogrammes (Fig. 6.1), le nombre de classes pour chaque groupe de variables.

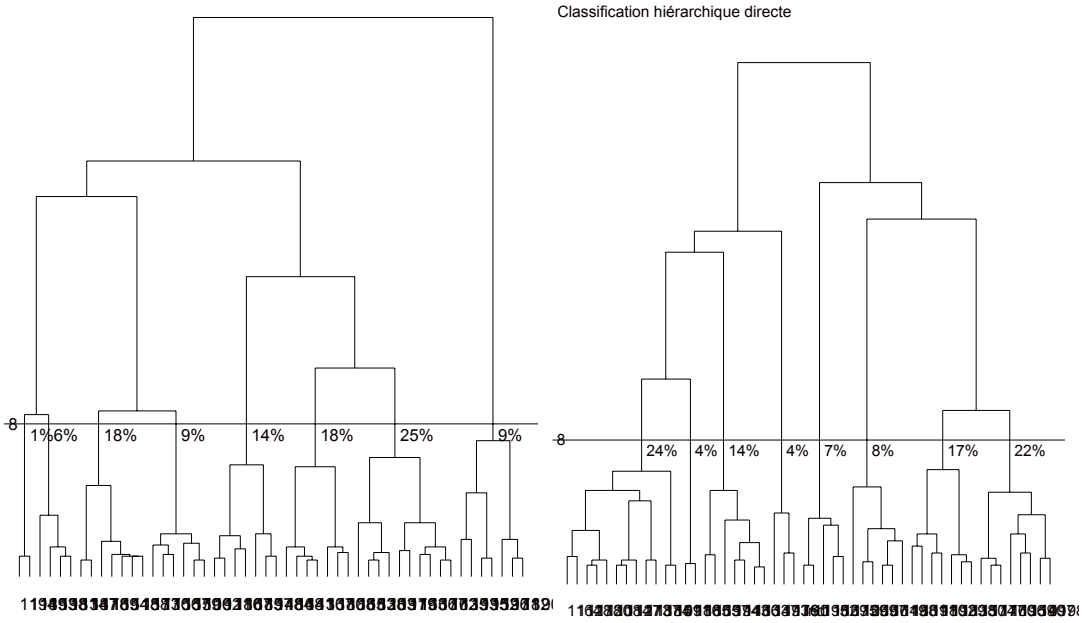


Fig.6.1 *Dendrogramme des deux groupes de variables(coupure en 8 classes)*

On choisit la coupure qui fournit une partition à 8 classes représentant pour le premier groupe de variables respectivement 1%, 6%, 18%, 9%, 14%, 18%, 25%, et 9% des sujets de l'échantillon, pour le deuxième groupe de variables respectivement 24%, 4%, 14%, 7%, 8%, 17% et 22%.

Les centres de ces 8 classes ont été reportés (Fig. 6.2) sur le meilleur plan factoriel issu de l'A.C.P., on a ainsi la position des classes les unes par rapport aux autres; La représentation des sujets sur ce plan permettrait de visualiser la dispersion de chacune des classes.

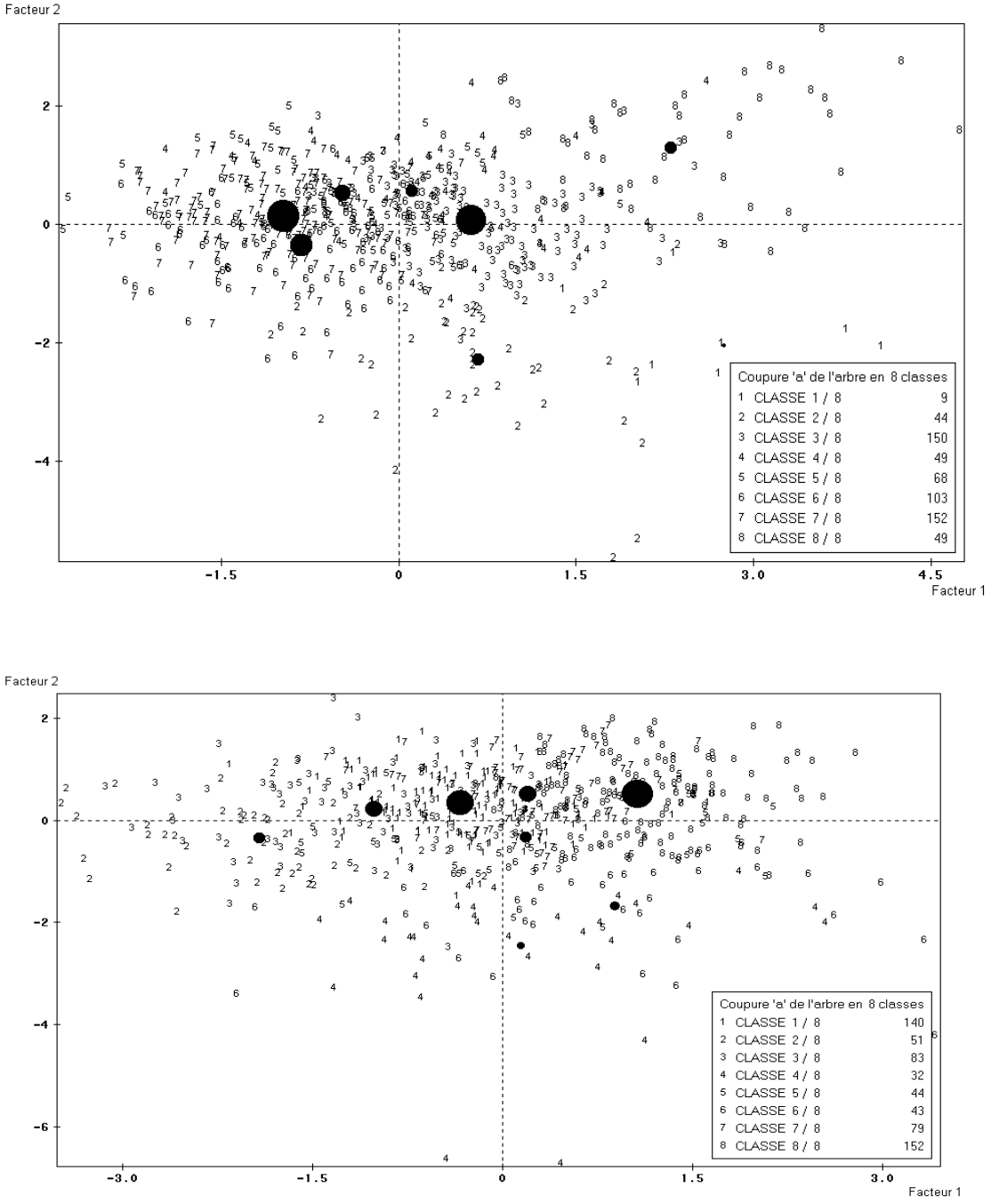


Fig.6.2 *Représentation des points- classes au barycentre des individus pour les deux groupes.*

On croise les deux partitions à 8 classes d'un même ensemble de données pour les deux groupes de variables, trouvées par la méthode de k-means, donnant les proportions

suivantes des 8 classes respectivement 24%, 7.5%, 16%, 15%, 13.5%, 3.5%, 14%, et 6.5% des sujets de l'échantillon.

On calcule les indices de ressemblance pour pouvoir les comparer. L'indice de Rand R' vaut 0.729, cette valeur proche de 1 ne suffit pas pour dire que les deux partitions sont proches, en effet cet indice donne la même importance aux couples d'individus qui sont ou non dans la même classe (accord global). On cherche l'indice dérivé de Jaccard, il prend la valeur 0.0979, celui de Janson et Vegelius vaut 0.035. On remarque que ces deux dernières valeurs sont faibles par rapport à l'indice de Rand R', cela revient à l'accord positif entre les deux partitions (Tab. 6.1).

$P_1 \setminus P_2$	**Même classe**	**Classes Différentes**
Même classe	5728	28830
Classes Différentes	23924	135894

$$\frac{n(n-1)}{2} = 194376$$

Tab 6.1 *Tableau d'accord croisant les deux partitions P_1 et P_2*

Afin de savoir si les partitions sont proches, nous appliquons notre méthodologie de profils latents pour étudier la distribution des indices de ressemblances. On utilise le logiciel LatentGOLD pour estimer un modèle de profils latents en 8 classes à partir de la totalité des données formées de 624 individus et 14 variables. Il nous donne les proportions des classes latentes, les moyennes et les variances, conditionnellement à chaque classe des variables qui doivent être indépendantes et de mêmes matrices de variances. On utilise ces paramètres pour simuler des échantillons de 624 individus à variables obéissant à ce modèle. On les coupe en deux ensembles de 7 variables, on cherche les partitions à 8 classes par la méthode des k-means puis on calcule les indices de Rand, de Jaccard, et de Janson Vegelius. On itère 100 fois. On obtient les distributions des ces indices dans la figure suivante (Fig. 6.3):

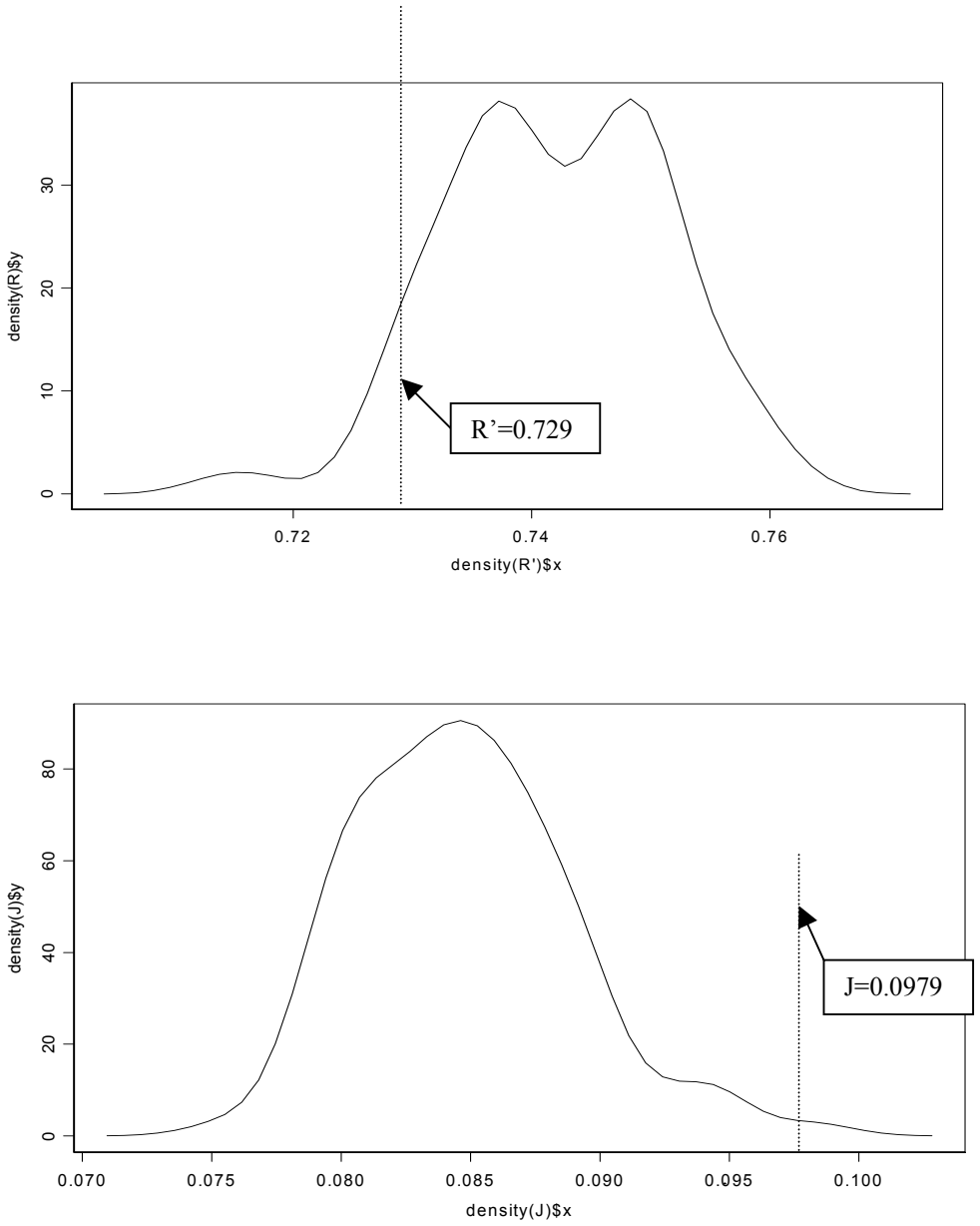


Fig.6.3 *Distribution de Rand et de Jaccard selon le modèle de profils latents*

En reportant les valeurs observées des indices sur leurs distributions, on remarque que la valeur de R' observée est plus petit de la moyenne de la distribution de R' trouvée par simulation qui est égale à 0.74203, mais la valeur observée de l'indice dérivé de Jaccard est suffisamment grand par rapport la moyenne de la distribution de cet indice qui vaut 0.08473678. Ce cas satisfaisant de l'indice d'accord positif nous permet de conclure que les partitions sont proches.

Le tableau de contingence croisant P_1 et P_2 est le suivant :

P_1 | P_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	ColTotl
1	2	9	7	6	9	5	15	4	57
2	42	13	21	20	22	4	18	6	146
3	2	0	1	2	3	1	3	1	13
4	39	5	24	12	9	5	10	11	115
5	16	3	6	14	3	0	11	5	58
6	5	2	2	7	6	0	9	5	36
7	35	7	32	26	26	6	21	7	160
8	7	8	8	6	6	1	1	2	39
ColTotl	148	47	101	93	84	22	88	41	624

Tab. 6.2 *Tableau de contingence initiale à P_1 et P_2*

Maintenant pour comparer les classes on compare leurs caractérisations selon les variables illustratives commune aux deux groupements. Mais il faut identifier les classes des deux partitions, alors on utilise le maximum du coefficient kappa afin de trouver la bonne numérotation des classes, en effet on permute les classes de la partition P_2 dans le tableau de contingence de base selon la valeur maximum de kappa. On réordonne les colonnes de ce tableau pour obtenir la valeur maximale de kappa, on trouve une valeur de 0.077032 pour la permutation de colonnes suivantes : 7, 5, 6, 1, 4, 8, 3, 2. Le tableau réordonné est alors le suivant :

P_1 | P_2	**7**	**5**	**6**	**1**	**4**	**8**	**3**	**2**	ColTotl
1	15	9	5	2	6	4	7	9	57
2	18	22	4	42	20	6	21	13	146
3	3	3	1	2	2	1	1	0	13
4	10	9	5	39	12	11	24	5	115
5	11	3	0	16	14	5	6	3	58
6	9	6	0	5	7	5	2	2	36
7	21	26	6	35	26	7	32	7	160
8	1	6	1	7	6	2	8	8	39
ColTotl	88	84	22	148	93	41	101	47	624

Tab. 6.3 *Tableau de contingence réordonné selon valeur maximale de kappa κ*

On remarque que la classe 1 de P_1 est identifiée à la classe 7 de P_2, la classe 2 de P_1 est identifiée à la classe 5 de P_2, la classe 3 de P_1 est identifiée à la classe 6 de P_2, la classe 4 de P_1 est identifiée à la classe 1 de P_2, la classe 5 de P_1 est identifiée à la classe 4 de P2, la classe 6 de P_1 est identifiée à la classe 8 de P_2, la classe 7 de P_1 est identifiée à la classe 3 de P_2, et la classe 8 de P_1 est identifiée à la classe 2 de P_2. On utilise les numérotations trouvées ci-dessus pour pouvoir comparer les descriptions des classes des partitions.

Pour tester la stabilité des classes, on utilise le test de Mc Nemar généralisé qui étudie la variation de proportions des classes pour les deux partitions. On obtient une valeur de T égale à 93.5394. Pour un risque de 5 %, cette valeur dépasse de loin le quantile de la loi de khi-deux de degré de liberté 28 qui est de 41.337. Ceci permet de rejeter l'hypothèse nulle H_0 et de conclure que les proportions des classes ont changé dans les deux partitions.

Pour trouver l'ordre des classes on peut aussi utiliser l'Analyse Factorielle des Correspondances, on permute les modalités selon leur classement sur le premier axe. En appliquant cette méthode aux partitions P_1 et P_2 du tableau (Tab. 6.2) et par utilisation du logiciel SPAD, le tableau réordonné est:

P_1 | P_2	**2**	**7**	**6**	**5**	**8**	**4**	**3**	**1**	ColTotl
1	9	15	5	9	4	6	7	2	57
3	0	3	1	3	1	2	1	2	13
6	2	9	0	6	5	7	2	5	36
8	8	1	1	6	2	6	8	7	39
7	7	21	6	26	7	26	32	35	160
5	3	11	0	3	5	14	6	16	58
2	13	18	4	22	6	20	21	42	146
4	5	10	5	9	11	12	24	39	115
ColTotl	47	88	22	84	41	93	101	148	624

Tab. 6.4 *Ordre selon le premier facteur en AFC*

En calculant le coefficient kappa pour ce tableau permuté on trouve la valeur de 0.03009489: La renumérotation selon le premier axe de l'AFC n'est donc pas optimale.

Comparaison des descriptions des classes

On compare les descriptions des classes de deux partitions après qu'on utilise les numérotations trouvées par la maximisation du coefficient kappa. Cette comparaison a été faite à partir des variables illustratives commune aux deux groupements.

La caractérisation des classes issues de la classification des données des deux groupes de variables est présentée dans le tableau 6.5 et 6.6, par exécution de la méthode PARTI /DECLA dans SPAD. Les colonnes CLA/MOD, MOD/CLA et GLOBAL fournissent respectivement le pourcentage de sujets présentant la modalité qui appartient à la classe, le pourcentage de la classe possédant la modalité et le pourcentage de sujets de l'échantillon global ayant la modalité, pour les partitions des deux groupes de variables.

CLASSE 1 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				1.44	CLASSE 1 / 8		aa1a	9
2.98	0.001	2.96	100.00	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
2.42	0.008	4.90	55.56	16.35	ouvrier qualifié	profession	cs03	102
2.05	0.020	2.54	88.89	50.48	Masculin	sexe	m	315
0.23	0.408	1.65	77.78	67.95	très importante	la preservation de l'environnement est une chose	env1	424
0.20	0.420	2.06	22.22	15.54	cadre moyen	profession	cs11	97
0.11	0.455	1.92	11.11	8.33	cadre supérieur	profession	cs14	52
0.04	0.484	2.27	11.11	7.05	autre employé qual.	profession	cs05	44
-0.06	0.477	0.00	0.00	7.85	ouvrier spécialisé	profession	cs02	49
-0.12	0.451	0.00	0.00	6.41	personnel de service	profession	cs07	40
-0.15	0.440	0.00	0.00	8.65	autre emp. non qual.	profession	cs06	54
-0.47	0.320	0.00	0.00	4.17	employé de commerce	profession	cs04	26
-0.47	0.319	0.00	0.00	11.86	non-réponse	profession	cs**	74
-0.50	0.310	0.00	0.00	4.01	petit commercant	profession	cs10	25
-0.69	0.244	0.00	0.00	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.75	0.228	0.57	11.11	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-2.05	0.020	0.32	11.11	49.52	feminin	sexe	f	309
-2.95	0.002	0.00	0.00	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317

CLASSE 2 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				7.05	CLASSE 2 / 8		aa2a	44
2.90	0.002	19.23	22.73	8.33	cadre supérieur	profession	cs14	52
2.76	0.003	16.22	27.27	11.86	non-réponse	profession	cs**	74
2.54	0.006	9.87	68.18	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
1.22	0.112	8.02	77.27	67.95	très importante	la preservation de l'environnement est une chose	env1	424
0.40	0.344	7.62	54.55	50.48	Masculin	sexe	m	315
0.32	0.373	8.25	18.18	15.54	cadre moyen	profession	cs11	97
0.10	0.461	8.00	4.55	4.01	petit commercant	profession	cs10	25
-0.15	0.439	3.85	2.27	4.17	employé de commerce	profession	cs04	26
-0.27	0.393	5.26	2.27	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.40	0.344	6.47	45.45	49.52	feminin	sexe	f	309
-0.69	0.244	3.70	4.55	8.65	autre emp. non qual.	profession	cs06	54
-0.98	0.163	2.27	2.27	7.05	autre employé qual.	profession	cs05	44
-1.01	0.156	5.11	20.45	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-1.66	0.049	0.00	0.00	6.41	personnel de service	profession	cs07	40
-1.98	0.024	0.00	0.00	7.85	ouvrier spécialisé	profession	cs02	49
-2.17	0.015	1.96	4.55	16.35	ouvrier qualifié	profession	cs03	102
-2.47	0.007	4.42	31.82	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317

CLASSE 3 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				24.04	CLASSE 3 / 8		aa3a	150
5.17	0.000	33.22	67.33	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
1.79	0.036	35.19	12.67	8.65	autre emp. non qual.	profession	cs06	54
1.75	0.040	36.36	10.67	7.05	autre employé qual.	profession	cs05	44
0.72	0.237	25.00	70.67	67.95	très importante	la preservation de l'environnement est une chose	env1	424
0.42	0.339	24.92	51.33	49.52	feminin	sexe	f	309
0.15	0.439	26.92	4.67	4.17	employé de commerce	profession	cs04	26
0.07	0.472	24.49	8.00	7.85	ouvrier spécialisé	profession	cs02	49
0.06	0.475	24.74	16.00	15.54	cadre moyen	profession	cs11	97
-0.01	0.494	22.50	6.00	6.41	personnel de service	profession	cs07	40
-0.36	0.361	21.62	10.67	11.86	non-réponse	profession	cs**	74
-0.37	0.356	22.73	26.67	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-0.42	0.339	23.17	48.67	50.48	Masculin	sexe	m	315
-0.50	0.308	21.57	14.67	16.35	ouvrier qualifié	profession	cs03	102
-0.55	0.292	15.79	2.00	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-1.78	0.037	8.00	1.33	4.01	petit commercant	profession	cs10	25
-5.23	0.000	15.14	32.00	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317

CLASSE 4 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				7.85	CLASSE 4 / 8		aa4a	49

1.55	0.061	12.37	24.49	15.54	cadre moyen	profession	cs11 97
1.38	0.084	11.76	24.49	16.35	ouvrier qualifié	profession	cs03 102
1.08	0.140	9.21	57.14	48.72	non	la famille est l'endroit ou on sent bien	fbi2 304
0.93	0.176	12.24	12.24	7.85	ouvrier spécialisé	profession	cs02 49
0.69	0.244	8.49	73.47	67.95	très importante	la preservation de l'environnement est une chose	env1 424
0.67	0.253	8.74	55.10	49.52	feminin	sexe	f 309
0.43	0.335	11.54	6.12	4.17	employé de commerce	profession	cs04 26
-0.09	0.463	6.76	10.20	11.86	non-réponse	profession	cs** 74
-0.13	0.448	5.26	2.04	3.04	peu importante	la preservation de l'environnement est une chose	env3 19
-0.32	0.375	5.00	4.08	6.41	personnel de service	profession	cs07 40
-0.34	0.368	5.56	6.12	8.65	autre emp. non qual.	profession	cs06 54
-0.50	0.308	4.55	4.08	7.05	autre employé qual.	profession	cs05 44
-0.67	0.253	6.98	44.90	50.48	Masculin	sexe	m 315
-0.76	0.224	6.25	22.45	28.21	assez importante	la preservation de l'environnement est une chose	env2 176
-0.84	0.201	3.85	4.08	8.33	cadre supérieur	profession	cs14 52
-1.15	0.124	0.00	0.00	4.01	petit commercant	profession	cs10 25
-1.31	0.095	6.31	40.82	50.80	oui	la famille est l'endroit ou on sent bien	fbi1 317

CLASSE 5 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				10.90	CLASSE 5 / 8		aa5a	68
4.18	0.000	16.09	75.00	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
1.15	0.125	20.00	7.35	4.01	petit commercant	profession	cs10	25
1.07	0.142	19.23	7.35	4.17	employé de commerce	profession	cs04	26
1.07	0.142	21.05	5.88	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
0.67	0.251	12.50	32.35	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
0.64	0.262	15.00	8.82	6.41	personnel de service	profession	cs07	40
0.59	0.277	14.29	10.29	7.85	ouvrier spécialisé	profession	cs02	49
0.40	0.343	13.64	8.82	7.05	autre employé qual.	profession	cs05	44
0.04	0.482	11.00	50.00	49.52	feminin	sexe	f	309
-0.04	0.482	10.79	50.00	50.48	Masculin	sexe	m	315
-0.13	0.450	9.26	7.35	8.65	autre emp. non qual.	profession	cs06	54
-0.16	0.435	10.78	16.18	16.35	ouvrier qualifié	profession	cs03	102
-0.35	0.363	9.28	13.24	15.54	cadre moyen	profession	cs11	97
-1.47	0.070	5.41	5.88	11.86	non-réponse	profession	cs**	74
-1.56	0.060	9.43	58.82	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-2.17	0.015	1.92	1.47	8.33	cadre supérieur	profession	cs14	52
-4.10	0.000	5.59	25.00	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304

CLASSE 6 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				16.51	CLASSE 6 / 8		aa6a	103
2.39	0.008	29.63	15.53	8.65	autre emp. non qual.	profession	cs06	54
1.55	0.061	18.93	58.25	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
1.09	0.137	21.62	15.53	11.86	non-réponse	profession	cs**	74
1.07	0.143	19.32	33.01	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
0.20	0.420	18.37	8.74	7.85	ouvrier spécialisé	profession	cs02	49
0.11	0.457	16.51	50.49	50.48	Masculin	sexe	m	315
0.00	0.499	17.50	6.80	6.41	personnel de service	profession	cs07	40
0.00	0.498	15.38	7.77	8.33	cadre supérieur	profession	cs14	52
-0.11	0.457	16.50	49.51	49.52	feminin	sexe	f	309
-0.14	0.444	15.91	6.80	7.05	autre employé qual.	profession	cs05	44
-0.29	0.386	12.00	2.91	4.01	petit commercant	profession	cs10	25
-0.29	0.385	15.79	2.91	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.37	0.355	14.71	14.56	16.35	ouvrier qualifié	profession	cs03	102
-0.81	0.209	15.57	64.08	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-1.37	0.086	11.34	10.68	15.54	cadre moyen	profession	cs11	97
-1.44	0.075	14.14	41.75	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
-2.40	0.008	0.00	0.00	4.17	employé de commerce	profession	cs04	26

CLASSE 7 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				24.36	CLASSE 7 / 8		aa7a	152
6.69	0.000	35.65	74.34	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
1.58	0.057	40.00	6.58	4.01	petit commercant	profession	cs10	25
1.45	0.074	38.46	6.58	4.17	employé de commerce	profession	cs04	26
1.41	0.080	35.00	9.21	6.41	personnel de service	profession	cs07	40
1.37	0.086	28.41	32.89	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
0.79	0.215	25.89	52.63	49.52	feminin	sexe	f	309
0.30	0.381	26.92	9.21	8.33	cadre supérieur	profession	cs14	52
0.02	0.490	26.32	3.29	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.07	0.471	23.53	15.79	16.35	ouvrier qualifié	profession	cs03	102
-0.42	0.337	20.45	5.92	7.05	autre employé qual.	profession	cs05	44
-0.54	0.296	21.65	13.82	15.54	cadre moyen	profession	cs11	97
-0.79	0.215	22.86	47.37	50.48	Masculin	sexe	m	315
-0.84	0.201	18.37	5.92	7.85	ouvrier spécialisé	profession	cs02	49
-1.02	0.154	18.92	9.21	11.86	non-réponse	profession	cs**	74
-1.15	0.124	22.88	63.82	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-1.59	0.056	14.81	5.26	8.65	autre emp. non qual.	profession	cs06	54
-6.75	0.000	12.50	25.00	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304

CLASSE 8 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				7.85	CLASSE 8 / 8		aa8a	49
4.14	0.000	12.50	77.55	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
1.42	0.078	9.52	61.22	50.48	Masculin	sexe	m	315
1.36	0.087	8.96	77.55	67.95	très importante	la preservation de l'environnement est une chose	env1	424
1.00	0.157	10.78	22.45	16.35	ouvrier qualifié	profession	cs03	102
0.93	0.176	12.24	12.24	7.85	ouvrier spécialisé	profession	cs02	49
0.79	0.215	10.31	20.41	15.54	cadre moyen	profession	cs11	97
0.49	0.312	12.00	6.12	4.01	petit commercant	profession	cs10	25
0.36	0.359	9.46	14.29	11.86	non-réponse	profession	cs**	74
0.13	0.448	10.53	4.08	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.25	0.400	5.77	6.12	8.33	cadre supérieur	profession	cs14	52
-0.32	0.375	5.00	4.08	6.41	personnel de service	profession	cs07	40
-0.50	0.308	4.55	4.08	7.05	autre employé qual.	profession	cs05	44
-1.21	0.114	0.00	0.00	4.17	employé de commerce	profession	cs04	26
-1.42	0.078	6.15	38.78	49.52	feminin	sexe	f	309
-1.46	0.073	5.11	18.37	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-1.56	0.059	1.85	2.04	8.65	autre emp. non qual.	profession	cs06	54
-4.07	0.000	3.47	22.45	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317

Tab. 6.5 *Caractérisation des classes par les modalités des variables illustratives de partition du premier groupe*

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES
DE Coupure 'a' de l'arbre en 8 classes
CLASSE 1 / 8

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				22.44	CLASSE 1 / 8		aa1a	140
2.73	0.003	38.89	15.00	8.65	autre emp. non qual.	profession	cs06	54
2.23	0.013	36.73	12.86	7.85	ouvrier spécialisé	profession	cs02	49
1.90	0.029	27.84	35.00	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
1.31	0.095	24.76	55.71	50.48	Masculin	sexe	m	315
1.26	0.103	34.62	6.43	4.17	employé de commerce	profession	cs04	26
0.72	0.237	31.58	4.29	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
0.62	0.268	27.50	7.86	6.41	personnel de service	profession	cs07	40
0.43	0.333	24.51	17.86	16.35	ouvrier qualifié	profession	cs03	102
0.26	0.396	23.03	52.14	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
0.11	0.456	22.73	7.14	7.05	autre employé qual.	profession	cs05	44
0.01	0.496	24.00	4.29	4.01	petit commercant	profession	cs10	25
-0.33	0.372	21.71	47.14	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
-1.11	0.134	15.38	5.71	8.33	cadre supérieur	profession	cs14	52
-1.14	0.128	17.53	12.14	15.54	cadre moyen	profession	cs11	97
-1.31	0.095	20.06	44.29	49.52	feminin	sexe	f	309
-2.37	0.009	19.58	59.29	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-3.25	0.001	8.11	4.29	11.86	non-réponse	profession	cs**	74

CLASSE 2 / 8

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				8.17	CLASSE 2 / 8		aa2a	51
4.70	0.000	13.49	80.39	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
1.77	0.038	13.40	25.49	15.54	cadre moyen	profession	cs11	97
1.21	0.112	9.20	76.47	67.95	très importante	la preservation de l'environnement est une chose	env1	424
1.08	0.139	13.64	11.76	7.05	autre employé qual.	profession	cs05	44
0.66	0.256	9.06	54.90	49.52	feminin	sexe	f	309
-0.08	0.470	5.26	1.96	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.19	0.423	7.69	7.84	8.33	cadre supérieur	profession	cs14	52
-0.25	0.401	8.11	11.76	11.86	non-réponse	profession	cs**	74
-0.37	0.355	3.85	1.96	4.17	employé de commerce	profession	cs04	26
-0.40	0.346	5.00	3.92	6.41	personnel de service	profession	cs07	40
-0.43	0.335	8.00	3.92	4.01	petit commercant	profession	cs10	25
-0.66	0.256	7.30	45.10	50.48	Masculin	sexe	m	315
-0.71	0.240	5.88	11.76	16.35	ouvrier qualifié	profession	cs03	102
-0.80	0.213	4.08	3.92	7.85	ouvrier spécialisé	profession	cs02	49
-0.93	0.175	6.25	21.57	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-1.00	0.159	3.70	3.92	8.65	autre emp. non qual.	profession	cs06	54
-4.63	0.000	3.15	19.61	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317

CLASSE 3 / 8

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				13.30	CLASSE 3 / 8		aa3a	83
2.44	0.007	26.53	15.66	7.85	ouvrier spécialisé	profession	cs02	49
2.13	0.017	20.59	25.30	16.35	ouvrier qualifié	profession	cs03	102
1.04	0.150	14.39	73.49	67.95	très importante	la preservation de l'environnement est une chose	env1	424
0.57	0.286	14.24	53.01	49.52	feminin	sexe	f	309
0.49	0.313	14.14	51.81	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
0.34	0.366	15.91	8.43	7.05	autre employé qual.	profession	cs05	44
-0.09	0.463	11.54	3.61	4.17	employé de commerce	profession	cs04	26
-0.17	0.431	12.00	3.61	4.01	petit commercant	profession	cs10	25
-0.25	0.403	11.11	7.23	8.65	autre emp. non qual.	profession	cs06	54
-0.35	0.364	10.00	4.82	6.41	personnel de service	profession	cs07	40
-0.39	0.347	12.62	48.19	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
-0.49	0.312	11.93	25.30	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
-0.57	0.286	12.38	46.99	50.48	Masculin	sexe	m	315
-0.66	0.255	5.26	1.20	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.77	0.220	10.31	12.05	15.54	cadre moyen	profession	cs11	97
-0.85	0.199	9.46	8.43	11.86	non-réponse	profession	cs**	74
-2.06	0.020	3.85	2.41	8.33	cadre supérieur	profession	cs14	52

CLASSE 4 / 8

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				5.13	CLASSE 4 / 8		aa4a	32
2.33	0.010	7.30	71.88	50.48	Masculin	sexe	m	315
1.38	0.083	7.39	40.63	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
1.11	0.134	7.84	25.00	16.35	ouvrier qualifié	profession	cs03	102
0.97	0.167	8.11	18.75	11.86	non-réponse	profession	cs**	74
0.35	0.362	5.26	3.13	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
0.33	0.370	5.59	53.13	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
0.28	0.390	7.69	6.25	4.17	employé de commerce	profession	cs04	26
-0.02	0.491	3.85	6.25	8.33	cadre supérieur	profession	cs14	52
-0.09	0.464	3.70	6.25	8.65	autre emp. non qual.	profession	cs06	54
-0.26	0.396	4.55	6.25	7.05	autre employé qual.	profession	cs05	44
-0.27	0.392	4.73	46.88	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
-0.42	0.337	5.00	6.25	6.41	personnel de service	profession	cs07	40
-0.64	0.261	0.00	0.00	4.01	petit commercant	profession	cs10	25
-0.71	0.238	3.09	9.38	15.54	cadre moyen	profession	cs11	97
-1.25	0.105	4.25	56.25	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-1.49	0.068	0.00	0.00	7.85	ouvrier spécialisé	profession	cs02	49
-2.33	0.010	2.91	28.13	49.52	feminin	sexe	f	309

CLASSE 5 / 8

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				7.05	CLASSE 5 / 8		aa5a	44
2.76	0.003	16.22	27.27	11.86	non-réponse	profession	cs**	74
2.11	0.017	9.39	65.91	49.52	feminin	sexe	f	309
1.92	0.028	12.37	27.27	15.54	cadre moyen	profession	cs11	97

1.90	0.029	9.21	63.64	48.72	non	la famille est l'endroit ou on sent bien	fbi2 304
1.57	0.058	8.25	79.55	67.95	très importante	la preservation de l'environnement est une chose	env1 424
1.53	0.063	13.46	15.91	8.33	cadre supérieur	profession	cs14 52
0.50	0.309	10.00	9.09	6.41	personnel de service	profession	cs07 40
-0.10	0.461	4.00	2.27	4.01	petit commercant	profession	cs10 25
-0.10	0.459	5.56	6.82	8.65	autre emp. non qual.	profession	cs06 54
-0.15	0.439	3.85	2.27	4.17	employé de commerce	profession	cs04 26
-0.27	0.393	5.26	2.27	3.04	peu importante	la preservation de l'environnement est une chose	env3 19
-1.38	0.083	4.55	18.18	28.21	assez importante	la preservation de l'environnement est une chose	env2 176
-1.81	0.035	0.00	0.00	7.05	autre employé qual.	profession	cs05 44
-1.84	0.033	5.05	36.36	50.80	oui	la famille est l'endroit ou on sent bien	fbi1 317
-1.98	0.024	0.00	0.00	7.85	ouvrier spécialisé	profession	cs02 49
-2.11	0.017	4.76	34.09	50.48	Masculin	sexe	m 315
-2.17	0.015	1.96	4.55	16.35	ouvrier qualifié	profession	cs03 102

CLASSE 6 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				6.89	CLASSE 6 / 8		aa6a	43
2.22	0.013	8.49	83.72	67.95	très importante	la preservation de l'environnement est une chose	env1	424
0.71	0.238	9.46	16.28	11.86	non-réponse	profession	cs**	74
0.65	0.257	8.82	20.93	16.35	ouvrier qualifié	profession	cs03	102
0.57	0.283	9.62	11.63	8.33	cadre supérieur	profession	cs14	52
0.57	0.286	7.62	55.81	50.48	Masculin	sexe	m	315
0.30	0.382	10.53	4.65	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
0.21	0.418	7.26	53.49	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
0.08	0.468	7.50	6.98	6.41	personnel de service	profession	cs07	40
0.07	0.474	8.00	4.65	4.01	petit commercant	profession	cs10	25
-0.04	0.486	6.19	13.95	15.54	cadre moyen	profession	cs11	97
-0.14	0.444	6.58	46.51	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
-0.46	0.324	4.08	4.65	7.85	ouvrier spécialisé	profession	cs02	49
-0.57	0.286	6.15	44.19	49.52	feminin	sexe	f	309
-0.65	0.259	3.70	4.65	8.65	autre emp. non qual.	profession	cs06	54
-0.94	0.173	2.27	2.33	7.05	autre employé qual.	profession	cs05	44
-1.04	0.150	0.00	0.00	4.17	employé de commerce	profession	cs04	26
-2.90	0.002	2.27	9.30	28.21	assez importante	la preservation de l'environnement est une chose	env2	176

CLASSE 7 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				12.66	CLASSE 7 / 8		aa7a	79
2.40	0.008	25.00	16.46	8.33	cadre supérieur	profession	cs14	52
0.81	0.210	16.22	15.19	11.86	non-réponse	profession	cs**	74
0.60	0.273	14.20	31.65	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
0.43	0.334	14.43	17.72	15.54	cadre moyen	profession	cs11	97
0.33	0.370	13.27	51.90	49.52	feminin	sexe	f	309
0.33	0.371	13.25	53.16	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
0.11	0.457	12.75	16.46	16.35	ouvrier qualifié	profession	cs03	102
0.02	0.493	13.64	7.59	7.05	autre employé qual.	profession	cs05	44
-0.15	0.440	10.53	2.53	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.19	0.423	11.54	3.80	4.17	employé de commerce	profession	cs04	26
-0.23	0.411	10.00	5.06	6.41	personnel de service	profession	cs07	40
-0.31	0.377	12.26	65.82	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-0.33	0.370	12.06	48.10	50.48	Masculin	sexe	m	315
-0.34	0.366	8.00	2.53	4.01	petit commercant	profession	cs10	25
-0.48	0.316	11.84	45.57	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304
-0.54	0.294	9.26	6.33	8.65	autre emp. non qual.	profession	cs06	54
-1.24	0.107	6.12	3.80	7.85	ouvrier spécialisé	profession	cs02	49

CLASSE 8 / 8

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				24.36	CLASSE 8 / 8		aa8a	152
3.80	0.000	30.91	64.47	50.80	oui	la famille est l'endroit ou on sent bien	fbi1	317
1.14	0.127	36.00	5.92	4.01	petit commercant	profession	cs10	25
0.34	0.365	25.57	29.61	28.21	assez importante	la preservation de l'environnement est une chose	env2	176
0.31	0.379	27.27	7.89	7.05	autre employé qual.	profession	cs05	44
0.23	0.409	24.92	50.66	49.52	feminin	sexe	f	309
0.11	0.454	26.92	4.61	4.17	employé de commerce	profession	cs04	26
0.06	0.475	25.00	6.58	6.41	personnel de service	profession	cs07	40
0.02	0.490	26.32	3.29	3.04	peu importante	la preservation de l'environnement est une chose	env3	19
-0.13	0.450	22.45	7.24	7.85	ouvrier spécialisé	profession	cs02	49
-0.14	0.445	24.07	8.55	8.65	autre emp. non qual.	profession	cs06	54
-0.16	0.438	24.32	11.84	11.86	non-réponse	profession	cs**	74
-0.23	0.409	23.81	49.34	50.48	Masculin	sexe	m	315
-0.28	0.391	22.68	14.47	15.54	cadre moyen	profession	cs11	97
-0.37	0.355	21.15	7.24	8.33	cadre supérieur	profession	cs14	52
-0.56	0.288	23.58	65.79	67.95	très importante	la preservation de l'environnement est une chose	env1	424
-1.63	0.052	17.65	11.84	16.35	ouvrier qualifié	profession	cs03	102
-3.86	0.000	17.43	34.87	48.72	non	la famille est l'endroit ou on sent bien	fbi2	304

Tab. 6.6 *Caractérisation des classes par les modalités des variables illustratives de partition du deuxième groupe.*

Description de la classe 1

Pour le premier groupe, la classe 1 représente 1.44% des modalités de l'échantillon global. Cette classe est caractérisée à 100% par des personnes qui ont l'opinion que la famille n'est pas l'endroit où on sent bien or on a 48.7% dans l'échantillon global et à

55% qui sont des ouvriers qualifiés. La modalité rare de la classe est la modalité femme qui a 11% (49.5% dans l'échantillon global).
Dans le deuxième groupe, cette classe représente 24.36% de l'échantillon global. Formées surtout des personnes répondant par oui la famille est l'endroit où on sent bien avec 74.34%. Les autres qui ont la réponse non à cette variable ne caractérisent pas beaucoup la classe à 25%.

Description de la classe 2

Pour le groupe 1, dans cette classe qui représente 7.05% de l'échantillon global, 22.7% des personnes sont des cadres supérieurs, 27.27% n'ont pas de réponse à la variable profession et 68.18% des individus de cette classe pensent que la famille n'est pas l'endroit où on sent bien alors qu'on compte 48.7% dans l'échantillon global. Les modalités qui caractérisent moins la classe sont « les ouvriers qualifiés » à 4.55% (16% dans l'échantillon global), et « la famille est l'endroit où on sent bien » avec 31.8% (58.8% dans l'échantillon global).

Cette classe dans la partition P_2 représente 10.9 % de l'échantillon global, et elle est caractérisée par 75% des personnes qui affirment que la famille est l'endroit où on sent bien contre 50.8% dans l'échantillon global. Elle est moins caractérisée par les autres personnes qui disent le contraire avec 25%.

Description de la classe 3

Cette classe représentant 24.04% de l'échantillon global dans la première partition, contient 67.33% des personnes répondant par non de la variable famille est l'endroit où on sent bien contre 48.72% dans l'échantillon global. La modalité oui de cette variable représente une modalité rare avec 32%.

Dans le groupe 2, la classe est caractérisée par les employées non qualifies avec 15.5% contre 8.65% dans l'échantillon global. Elle n'est pas caractérisée par des employées de commerces.

Description de la classe 4

A 7.85% de l'échantillon global, dans le groupe 1, cette classe a 24.49% des cadres moyennes contre 15.54% de l'échantillon global et des ouvriers qualifiés à 24.49% contre

16.35% dans l'échantillon global. Elle est rarement caractérisée par la modalité oui de la variable famille est l'endroit où on sent bien avec 40.82% contre 50.8% de l'échantillon global.

Pour le deuxième groupe, cette classe représente 22.44% de l'échantillon global, est caractérisée à 15% des personnes qui ont un emploi non qualifié (8.65% dans l'échantillon global), et 12.86% des ouvriers spécialisés. Les modalités rares sont les non-réponses de la variable profession avec 4.29% (11.86% dans l'échantillon global) et la préservation de l'environnement est une chose très importante avec 59.29% (67.95% dans l'échantillon global).

Description de la classe 5

Au premier groupe, la 5^{ieme} classe représente 7% de l'échantillon global. Les femmes et les personnes qui n'ont pas de réponses à la profession caractérisent cette classe par 65.9% et 27.27% respectivement. Les modalités rares sont les hommes à 34.09% contre 50.48% et les ouvriers qualifiés à 4.55% contre 16.35% dans l'échantillon global.

A 5% de l'échantillon global, elle est formée, dans la deuxième partition, par 71.88% des hommes contre 50.48% de l'échantillon global et rarement caractérisée par les femmes (28.1% contre 49.52% de l'échantillon global).

Description de la classe 6

A 6.89 % de l'échantillon global, pou le groupe 1, et formées à 83,7% des personnes qui pensent que la préservation de l'environnement est très important contre 67.95% dans l'échantillon global. Ceux qui ont répondu par peu important caractérisent rarement la classe avec 9.3%.

Dans le groupe 2, la classe représente 24.36% de l'échantillon global. Contrairement à la précédente elle est caractérisée par la modalité oui de la variable « la famille est l'endroit où on sent bien », avec 64.4%.

Description de la classe 7

La classe représente 12.66% de l'échantillon global dans la première partition, et elle est caractérisée à 16.46% par les personnes ayant une profession de cadres supérieurs contre

8.33% dans l'échantillon global. Les ouvriers spécialisés sont rares dans la classe à 3% contre 7.85% de l'échantillon global.

Dans la deuxième partition, cette classe représente 13.3% de l'échantillon global ayant 15.66% des ouvriers spécialisés et 25.3% ouvriers qualifiés. Ceux qui sont cadres supérieurs caractérisent peu cette classe avec 2.4% contre 8.33% de l'échantillon global.

Description de la classe 8

A 7.85% de l'échantillon global, la classe est caractérisée par 77.55% contre 48.72% dans l'échantillon global de la modalité non pour la variable « la famille est l'endroit où on sent bien », la modalité oui est une modalité rare avec 22.45%.

Dans le groupe 2, la classe 8 représente 8.17% de l'échantillon global, dont 80.39% pensent que la famille n'est pas l'endroit où on sent bien (48.7% dans l'échantillon global). On a 19% ont répondu le contraire ce qui lui fait une modalité rare. On peut alors conclure que cette modalité a la même répartition dans les deux partitions.

A partir des descriptions des 8 classes, on peut confirmer que les variables significatives d'une classe à l'exception de la classe 8 n'ont pas les mêmes répartitions pour les deux groupements de variables ayant les mêmes individus. Cela peut être logique comme ayant deux partitions qui ne sont pas très proches.

6.4 Comparaison de partitions de deux ensembles d'individus avec mêmes variables

6.4.1 Comparaison par projection des partitions

On applique la méthode de comparaison par projection des partitions présentée dans le paragraphe 5.3 du chapitre 5.

L'idée consiste à obtenir une distribution des valeurs de l'indice de Rand trouvées après 50 itérations en utilisant les étapes suivantes:

- Permutation aléatoire du fichier de données et coupure en moitié pour trouver les deux échantillons à 312 individus chacun.
- Recherche de la partition de l'un des échantillons par la méthode de k-means à 4 classes

- Projection des individus du deuxième échantillon sur la partition trouvée en utilisant la méthode discrim implantée dans SAS.
- Recherche d'une autre partition par la méthode des k-means à 4 classes de ce même deuxième échantillon.
- Croisement des numéros de classes des partitions trouvées par k-means et par projection.
- Calcul de l'indice de Rand.

On obtient la distribution de l'indice de Rand illustrée dans la figure (Fig. 6.4). Les valeurs de Rand pour cette distribution ont une moyenne de 0.665 et d'écart type égal à 0.03. la valeur la plus fréquente est autour de 0.67.

D'autre part, on partage le fichier de données selon la variable sexe et on traite par la même procédure pour obtenir les deux valeurs de Rand correspondantes aux deux modalités homme et femme. : pour les Hommes (315 individus) R_H= 0.6311 et pour les Femmes (309 individus) R_F=0.623. Ces deux valeurs sont reportées sur la distribution afin de les comparer à celles obtenues par découpage aléatoire des données.

On peut remarquer que (Fig. 6.4) les indices en projection 1 sur 2 ou 2 sur 1 sont proches, ce qui est satisfaisant puisque le problème est symétrique. La valeur est cependant trop faible pour pouvoir dire que la partition effectuée sur les hommes est proche de celle effectuée sur les femmes.

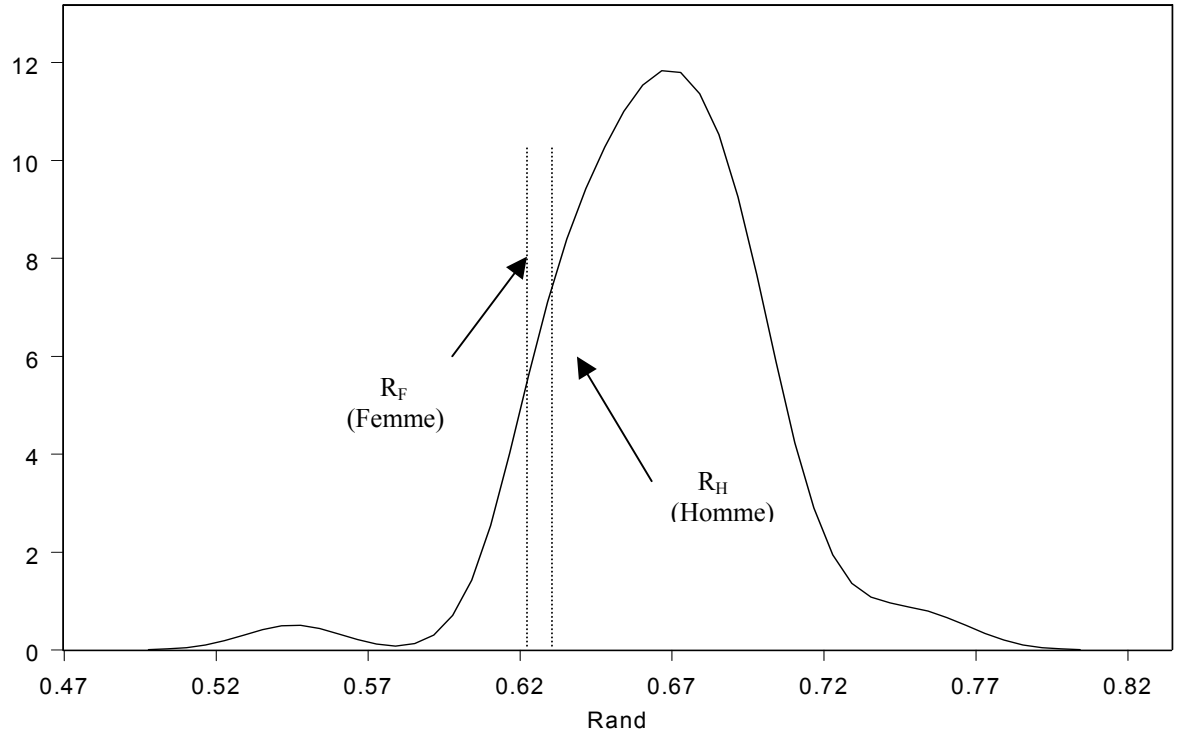


Fig.6.4 *Distribution de Rand après permutations aléatoires et projections*

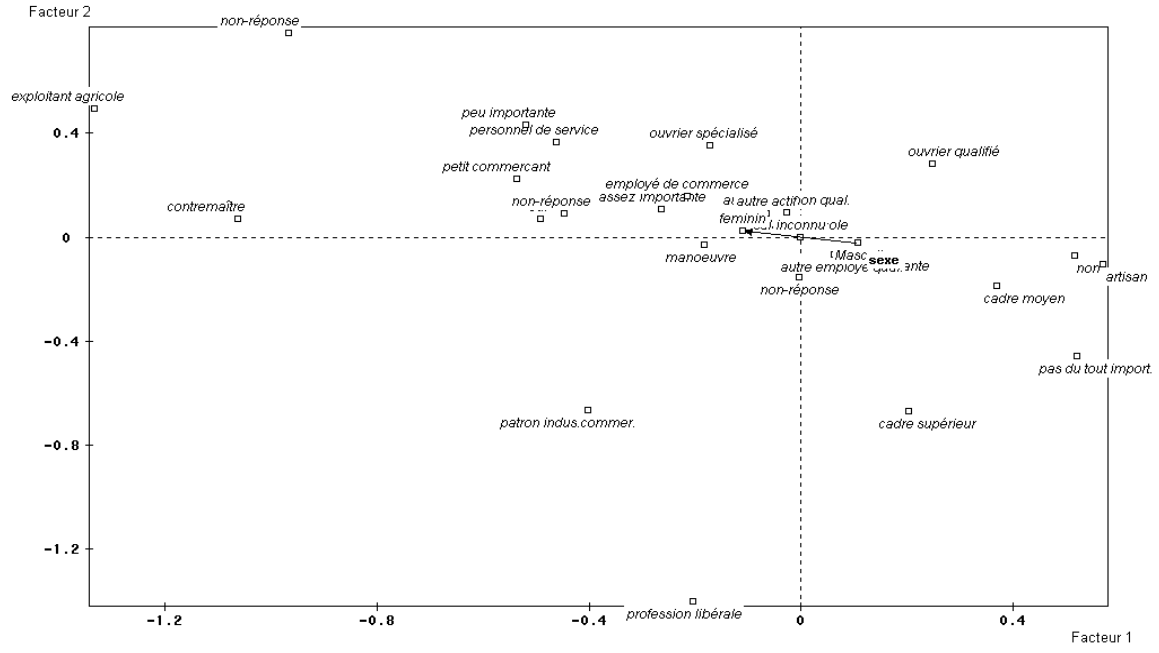


Fig. 6.5 *Représentation graphique des modalités des variables illustratives après un ACP et la trajectoire de la variable sexe.*

6.4.2 Comparaison des classifications de variables

Cette approche consiste à comparer les deux arbres hiérarchiques trouvés par une méthode de classification de variables. Pour chaque échantillon, on cherche la matrice de similarité en utilisant le coefficient de corrélation linéaire. On obtient les deux structures hiérarchiques par agrégation des variables. On calcule ensuite le coefficient de Spearman entre les deux ultramétriques.

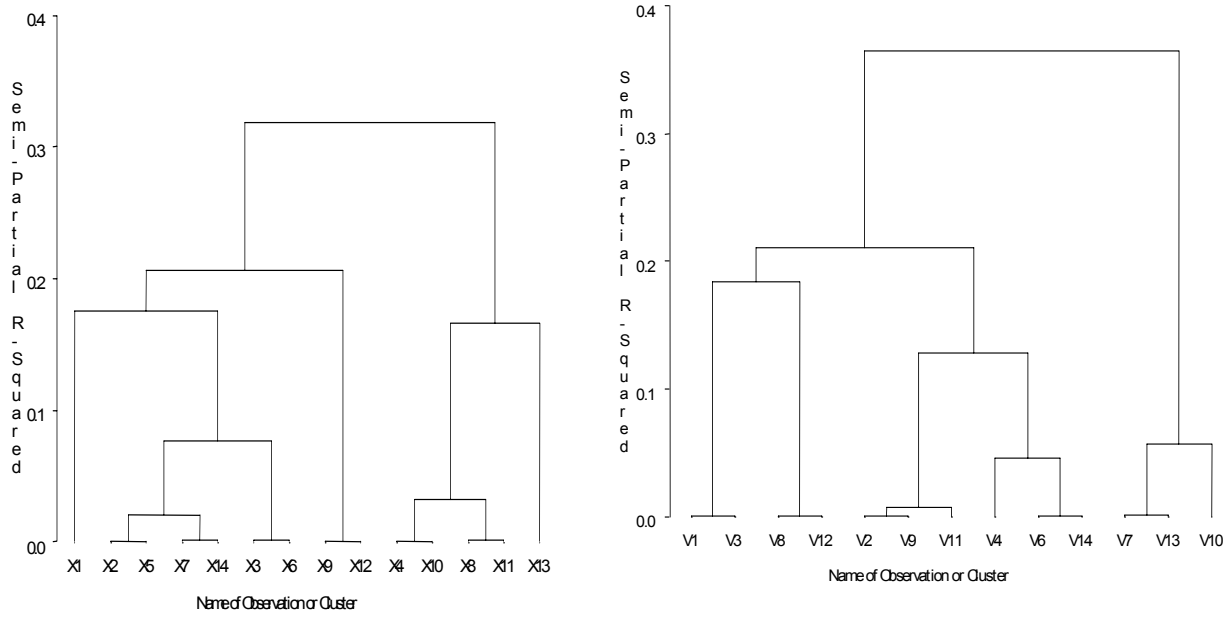


Fig.6.6 *Dendrogramme pour les échantillons selon homme et femme respectivement*

Le coefficient de Spearman prend la valeur–0.032. Pour un risque de 5%, la valeur critique r' est égal à 0.538. Il n'existe donc pas de corrélation entre les ultramétriques et on peut conclure que les structures des classifications hiérarchiques sont significativement différentes selon que l'on est homme ou femme.

Remarque : Dépasser le seuil à 5% (ce qui n'est pas le cas ici) n'est pas suffisant pour dire que les hiérarchies sont proches. Il faudrait pouvoir étudier la distribution du coefficient de Spearman entre ultramétriques quand les données sont issues du même modèle, selon la méthodologie développée au chapitre 5. Cette étude reste à faire.

Conclusion

Le travail que nous venons de présenter dans ce mémoire traite de la comparaison de structures de classifications données à travers l'étude des différents indices de ressemblances entre partitions. Il s'est articulé autour de deux axes principaux qui sont :

- Même ensemble d'individus avec deux groupes de variables
- Deux échantillons indépendants d'individus décrits par les mêmes variables

Le déroulement des étapes de notre travail et les résultats obtenus sont les suivants :

Dans le premier chapitre, un panorama servant à la compréhension des modèles probabilistes qui évaluent et étudient l'existence d'une partition, les algorithmes de classifications tels que les k-means et les algorithmes ascendants, et les travaux traitant les problématiques de la validation et de la détermination du « vrai » nombre des classes d'une partition, a été présenté.

Le deuxième chapitre a été consacré aux méthodes d'interprétation des classes d'une partition d'un ensemble de données. Ce chapitre aborde dans un premier temps les méthodes classiques utilisées en analyse de données basées sur les caractéristiques des individus appartenant à une même classe à partir des modalités des variables d'une partition. En deuxième lieu, nous avons évoqué les travaux s'appuyant sur l'analyse des données symboliques et offrant une aide à l'interprétation des résultats, au moyen de règles logiques tels que, la méthode CABRO proposé par H.T.Bao [BAO 88], la méthode proposée par M.Gettler- Summa [GET 93] appelée marquage sémantique, et les méthodes de classification divisives proposée par M.Chavent [CHA 97].

Le troisième chapitre étudie en détail les indices de comparaison de deux partitions. Nous avons examiné différents indices de comparaison: en plus de l'indice bien connu de Rand et celui corrigé par Hubert, L.[HUB 85], nous avons étudié sa version asymétrique [CHAV 01] utilisée pour la comparaison de partitions emboîtées, avec des nombres différents de classes. Nous avons ajouté deux autres indices inspirés de test de Mc Nemar et de l'indice de Jaccard. L'indice de corrélation vectorielle introduit par Robert, P. et Escoufier, Y. [ROB 76] qui s'est révélé identique au coefficient de Janson, S. et Vegelius, J.[JAN 82], le coefficient kappa de Cohen [COH 60], l'indice de redondance proposé par Stewart et Love [STE 68], ainsi que l'indice de Popping [POP 83], ont été présentés.

Le chapitre quatre développe notre méthodologie basée sur un modèle de profils latents pour comparer des partitions proches ayant des variables différentes pour un même ensemble d'individus. Une étude distributionnelle des différents indices de ressemblances a été effectuée. A base des simulations, l'effet des paramètres tels que la séparation des classes, le nombre d'individus, et le nombre de classes des partitions sur ces différents indices a été discuté. Les tests de stabilité d'une classification ou d'homogénéité ont été présentés.

Dans le chapitre cinq, nous évoquons les méthodes et les tests classiques de comparaison des partitions provenant des données de mêmes variables mais de différents individus. En se basant sur la projection des partitions, une nouvelle méthode de comparaisons a été proposée. Une autre approche pour la comparaison par utilisation de la classification des variables a été développée. Enfin, la stabilité des interprétations des classes des partitions a été présentée.

Le dernier chapitre a été consacré à quelques exemples de traitement de données réelles pour montrer la mise en œuvre et la pertinence des approches développées dans les chapitres précédents.

Perspectives

Dans la perspective de ces travaux, l'un des axes à développer est d'envisager l'utilisation des modèles probabilistes plus généraux autre que celui du modèle de profils

latents). Ces modèles plus généraux peuvent éventuellement incorporer des données mixtes qualitatives et quantitatives.

Un autre axe sera à considérer, celui qui concerne l'extension de ces travaux à la comparaison de plus de deux partitions ou hiérarchies. C'est le cas pratique lors du traitement d'enquêtes ou de panels avec T dates d'observation.

Enfin, il nous semble intéressant de continuer ces travaux dans le but de trouver une généralisation pour comparer des partitions « proches » ou de classifications hiérarchiques.

Bibliographie

[AKA 73] AKAIKE, H., Factor Analysis and AIC, *Psychometrika*, vol. 52, 317-332, 1973.

[BAC 85] BACELAR- NICOLAU, H., The Affinity coefficient in Cluster Analysis, in *Methods of Operation Research*, Martin J. Bekman et al. Ed., Verlag Anton Hain, Münich, vol. 53, 507-512, 1985.

[BAC 88] BACELAR- NICOLAU, H., Two Probabilistic Models for Classification of Variables in Frequency Tables. *Classification and Related Methods; H.H. Bock (Ed.), North Holland,* 181-189, 1988.

[BAC 02] BACELAR- NICOLAU, H., *On the Generalised Affinity Coefficient for Complex Data,* Byocybernetics and Biomedical Engineering, vol. 22 (1), 31-42, 2002.

[BAI 82] BAILEY, T.A., DUBES R., Cluster Validity Profiles, *Pattern Recognition*, vol. 15 (2), 61-83, 1982.

[BAO 88] BAO, H.T, DIDAY, E., GETTLER- SUMMA, M., Generating Rules for Expert System from Observation, *en Pattern Recognition Letters*, 265-271, 1988.

[BAO 91] BAO, H.T, HUYEN, T., A Method for Generating Rules from Examples and its Application. *Symbolic and Numeric Data Analysis and Learning, ed Diday,* Nova Science, 493-504, 1991.

[BAR 63] BARNARD, G.A., Discussion of a Paper by M.S. Barlett, *Journal of the Royal Statistical Society, Series B*, vol. 25, 294, 1963.

[BAR 99] BARTHOLOMEW, D.J., KNOTT, M., *Latent Variable Models and Factor Analysis* , Arnold, London,1999.

[BEL 98] BEL MUFTI G., *Validation d'une Classe par Estimation de sa Stabilité*, thèse de PhD, Université Paris IX Dauphine, octobre 1998.

[BEN 02] BEN –HUR, A., ELISSEFF, A., GUYON, I., A Stability Based Method for Discovering Structure in Clustered Data, *Pacific Symposium on Biocomputing,* Altman, R., Dunker, A., Hunter, L., et al. Eds., World Scientific, 6-17, 2002.

[BIE 00] BIERNACKI, C., CELEUX, G., GOVAERT, G., Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Trans., on PAMI*, vol. 22, 719-725, 2000.

[BOC 77] BOCK, H.H, On Tests Conserning the Existence of a Classification. *First International Symposium on Data Analysis and Informatics.* INRIA, Rocquencourt, 449-464, 1977.

[BOC 85] BOCK, H.H, On Some Significant Tests in Cluster Analysis, *Journal of Classification*, vol. 2, 77- 108, 1985.

[BOC 99] BOCK, H.H.The Classical Data Situation, *Analysis of Symbolic, in series : Studies in Classification, Data Analysis, and Knowledge* Organisation, Springer, 24- 39, 1999.

[BOC 00] BOCK, H.H., DIDAY, E. (eds.). Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, *Series: Studies in Classification, Data Analysis, and Knowledge* Organisation, , Springer, Berlin, vol. 15, 2000.

[BOZ 88] BOZDOGAN, H., ICOMP: A New Model Selection Criterion, *H H. Bock Ed., Classification and Related Methods of Data Analysis,* North- Holland, Amsterdam, 599-608, 1988.

[BOZ 94] BOZDOGAN, H., Mixture-model cluster Analysis using Model Selection Criteria and a New informational measure of complexity, *H. Bozdogan Eds., Multivariate statistical modeling*, vol. 2, 69-113, Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach. Dordrecht: Kluwer Academic, 1994.

[BOZ 00] BOZDOGAN, H., Akaïke's Information Criterion and Recent Developments in Information Complexity, *Journal of Mathematical Psychology,* vol. 44,62-91, 2000.

[BOZ 03] BOZDOGAN, H.,, Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms, *Proceeding of JISS 2003, Lisbonne*, vol. 2, 15-56, 2003.

[BRE 89] BRECKENRIDGE J.N., Replicating Cluster Analysis: Method, Consistency, and Validity, *Multivariate Behavior Research*, vol. 24 (2), 147-161, 1989.

[BRE 74] BRENNAN, R.L, LIGHT, R.J., Measuring Agreement when Two Observers Classify People into Categories not Defined in Advance*, British Journal of Mathematical and Statistical Psychology*, vol 27, 154-163, 1974.

[BRI 84] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. et STONE, C.J., *Classification and Regression Trees*. Wadsworth & Brooks/ code advanced book & software, 1984.

[BRI 91] BRITO, P., *Analyse de Données Symboliques Pyramides d'Héritage*, Thèse de PhD, Université Paris IX Dauphine, 1991.

[CEL 92] CELEUX, G., Résultats Asymptotiques et Validation en Classification. In *Modèles pour l'Analyse de Données Multidimensionnelles.* Droesbeke, J.J., Fichet, B., Tassi, P., Eds., Economica, Paris, 1992.

[CEL 94] CELEUX, G., NAKACHE, J.P., *Analyse Discriminante sur Variables Qualitatives*. Polytechnica Editions, 1994.

[CEL 96] CELEUX, G., An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model, *Journal of Classification*, vol. 13, 195-212, 1996.

[CHA 97] CHAVENT, M., *Analyse de Données Symboliques. Une méthode divisive de classification*- Thèse, Université Paris IX -Dauphine, 1997.

[CHA 99] CHAVENT, M., Criterion- Based Divisive Clustering for Symbolic Data, *Ed. Bock, H.H., et Diday, E., Analysis of Symbolic Data*, 299-311, 1999.

[CHA 01] CHAVENT, M., ET AL., Critère de Rand Asymétrique, *in Proceedings SFC 2001, 8èmes rencontres de la Société Francophone de Classification*, Pointe à Pitre, 2001.

[CHE 95] CHEESEMAN, P., STUTZ, J., *Bayesian Classification (Autoclass) : Theory and Result*, U.M. Fayyad, G.Piatetsky- Shapiro, P.Smyth and R. Uthurusamy (eds.), Advances in Knowledge Discovery and Data Mining, Menlo Park : the AAAI Press, 1995.

[COH 60] COHEN J., A Coefficient of Agreement for Nominal Scales*., Educ. Psychol. Meas*., vol. 20, 27-46, 1960.

[CON 80] CONOVER, W. J., *Practical Nonparametric Statistics*, 2^{e} édition New York: John Wiley & Sons, 1980.

[DAY 83] DAY, W.H.E, the Role of Complexity in Comparing Classifications, *Mathematical Biosciences*, vol. 66, 97- 114, 1983.

[DAY 98] Dayton, C.M, *Latent Class Scaling Analysis*, Series : Quantitative Applications in the Social Sciences, 126, SAGE publications, 1998.

[DEC 92] DECARVALHO, F.A.T, *Méthode Descriptive en Analyse de Données Symboliques*, thèse de PhD, Université Paris IX Dauphine, 1992.

[DEC 94] DE CARVALHO, F.A.T, Proximity Coefficients Between Boolean Symbolic Objects, Diday, E. et al.(eds.): *New Approaches in Classification and Data Analysis, Series: Studies in Classification, Data Analysis, and Knowledge Organisation,* vol. 5, Springer-Verlag, Berlin, 387-394, 1994.

[DEC 98] DE CARVALHO, F.A.T, Extension Based Proximity Coefficients Between Constrained Boolean Symbolic Objects. *Hayashi, C. et al. (eds.): Proc. of IFCS'96*, Springer, Berlin, 370-378, 1998.

[DEG 90] DEGLAS, M., Vers une Représentation Logique du Sens, *Rapport technique N° 24-90, Université Pierre et Marie Curie Paris 6*, 1990.

[DID 71] DIDAY, E., La Méthode de Nuées Dynamiques, *Revue de la Statistique Appliquée,* vol. 19(2), 19-34, 1971.

[DID 80] DIDAY, E., et al., Clustering in Pattern Recognition. Ed. J.C. Simon, Proc. NATO Advanced Study Institute on Digital Processing and Analysis. Bonas, France 1980.

[DID 91] DIDAY, E., Des Objets de l'Analyse de Données A Ceux de l'Analyse des Connaissances. *Inductions symboliques et numériques à partir des données- ed par Kodratoff, Y et Diday, E.,* Cépaduès, 1991.

[DID 92] DIDAY, E., Eléments d'Analyse des Données Symboliques. *PRC IA Apprentissage Symbolique et Numérique. Marseille*, décembre 1992.

[DID 94] DIDAY, E., LECHEVALLIER, Y., SCHADER, M., BERTRAND, P., et BURTSCHY, B. Ed., *New Approaches in Classification and Data Analysis*. Proc. Conf. De IFCS (IFCS-93), Springer-Verlag, Heidelberg, 1994.

[ELA 90] ELAYOUBI, N., *Liaison entre Analyse Factorielle et analyse Relationnelle*, Thèse de doctorat de l'Université de Paris 6, 1990.

[ESP 00a] ESPOSITO, F., MALERBA, D., TAMMA, V., et BOCK, H.H. Classical Resemblance Measures. *Bock, H.H, Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation*, Springer-Verlag, Berlin, vol. 15, 139-152, 2000.

[ESP 00b] ESPOSITO, F., MALERBA, D., TAMMA, V., Dissimilarity Measures for Symbolic Objects. *Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation*, Springer-Verlag, Berlin, vol. 15, 165-185, 2000.

[EVE 81] EVERITT, B., HAND, D.J., *Finite Mixture Distributions*, Chapman and Hall, London, 1981.

[EVE 93] EVERITT, B.S., *Cluster Analysis*, Edward Arnold, London, 1993.

[FOW 83] FOWLKES, E.B. AND MALLOWS, C.L., A Method for Comparing Two Hiearchical Clusterings, J*ournal of American Statistical Association*, vol. 78, 553- 569, 1983.

[FOR 65] FORGY, E.W., Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications. *Biometric Society Meetings*, Riverside, California (Abstract in: Biometrics Vol. 21, No 3, 768,), 1965.

[GET 94] GETTLER-SUMMA, M., PERINEL, E., et FERRARIS, J. Automatic Aid to Symbolic Cluster Interpretation*. In New Approaches in Classification and Data Analysis,* Diday, E. et al..Eds. IFCS-93, 405-413, 1994.

[GET 98] GETTLER- SUMMA, M., Approches MGS Marquage et Généralisation Symboliques pour de Nouvelles Aides a l'Interprétation en Analyse de Données. *Cahier du Cérémade n^0 9830 Université* Paris IX Dauphine, France, 1998.

[GET 00] GETTLER- SUMMA, M., Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis. *Ed. Kiers, H.A.L., Rasson J.P., Groenen, et al. : Proc. of IFCS'00*, Namur, Belgium, 2000.

[GIL 89] GILLICK, L., COX, S., Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *ICASSP 89*, vol.1, 532-535,1989.

[GNA 77] GNANADESIKAN R., KETTENRING J.R., LANDWEHR J.M., Interpreting and Assessing the Results of Cluster Analysis. *Bulletin of International Statistical Institute*, vol. 47 (2), 451-463, 1977.

[GOO 79] GOODMAN, L., KRUSKAL, W., *Measures of Association for Cross- Classifications,* Springer-Verlag, New York, 1979.

[GOR 87] GORDON, A.D., A Review of Hierarchical Classification*, J.R Statistics Soc., A,* vol. 150, Part2, 119-137, 1987.

[GOR 88] GORDON, A.D., CATA, A.De., Stability and Influence in Sum of Squares Clustering*, Metron*, vol. 46, 347-360, 1988.

[GOR 98] GORDON, A.D., *Cluster Validation*, Studies in Classification, Data Analysis, and Knowledge Organization : Data Science, Classification, and Related Methods, Ed. Hayashi C., Ohsumi N., Yajimi K., Tanaka Y., Bock H.H., Baba Y., 493-504, Springer-Verlag, 1998.

[GOW 94] GOWDA, K. C., DIDAY, E. Symbolic Clustering Using a New Dissimilarity Measure. *In Pattern Recognition,* vol. 24 (6), 567-578, 1994.

[HAL 01] HALKIDI, M., VAZIRGIANNIS, M., Clustering Validity Assessment : Finding the Optimal Partitioning of a Data Set, *Proceedings of ICDM Conference, California*, USA, 2001.

[HAN 81] HAND, D.J*., Discrimination and Classification*, Wiley, London, 1981.

[HEI 96] Heinen, T., *Latent Class and Discrete Latent Trait Models, Similarities and Differences*, Advanced Quantitative Technics in the Social Sciences, SAGE publications, 1996.

[HIL 98] HILLALI, Y., *Analyse et Modélisation des Données Probabilistes : Capacités et Lois Multidimensionnelles,* Thèse de PhD, université Paris IX Dauphine, 1998.

[HOP 68] HOPE A.C.A., A Simplified Monte Carlo Significance Test Procedure, *Journal of The Royal Statistical Society, Series B*, vol. 30, 582-598, 1968.

[HUB 85] HUBERT L., ARABIE P., Comparing Partitions, *Journal of Classification*, vol. 2, 193-198, 1985.

[ICH 94] ICHINO, M., YAGUCHI, H. Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, No. 4, 698-707, 1994.

[IDR 00] IDRISSI A., *Contribution à l'Unification de Critère d'Association pour Variables Qualitatives*, Thèse de doctorat de l'Université de Paris 6, 2000.

[IMH 61] IMHOF P. Computing The Distribution of Quadratic Forms in Normal Variables, *Biometrika,* vol. 48, 419- 426.

[JAI 87] JAIN, A.K., MOREAU, J.V., Bootstrap Technique in Cluster Analysis, Pattern Recognition, Vol. 20, 547-568, 1987.

[JAI 88] JAIN, A.K., DUBES, R., *Algorithms for Clustering Data*. Englewood Cliffs. NJ. Prentice-Hall,1988.

[JAN 82] JANSON S., VEGELIUS J. The J-index as a Measure of Association for Nominal Scale Response Agreement*. Applied psychological measurement*, vol. 16, 243-250, 1982.

[JAR 71] JARDINE, N. et SIBSON, R*., Mathematical Taxonomy*. Londres, Willey Ed., 1971.

[KEN 61] KENDALL M.G, STUART A. The Advanced Theory of Statistics*, Griffin, Londre*, vol. 2, 1961.

[KOE 69] KOERTS J., ABRAHAMSE A.P.J., On The Theory and Application of The General Linear Model, *Rotterdam University Press,* Rotterdam, 1969.

[KOD 91] KODRATOFF, Y., DIDAY, E. Ed., *Introduction Symbolique et Numérique à Partir de Données*. CEPADUES, Toulouse, 1991.

[KRI 99] KRIEGER A., GREEN P., A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base*, Journal of Classification*, vol. 16, 63-89,1999.

[LAN 67] LANCE, G.N., WILLIAMS, W.T., A General Theory of Classification Sorting Strategies, *Computer Journal,* vol.9, 373-380, 1967.

[LAZ 50] LAZARSFELD, P.F , *The Logical and Mathematical Foundation of Latent Structure Analysis*, In S. Stouffer (Ed.), Measurement and Prediction, 362-412, Princeton, N.J :Princeton University Press,1950.

[LAZ 68] LAZARSFELD, P.F., HENRI, N.W., *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.

[LAZ 01] LAZRAQ, A., CLEROUX R.. Statistical Inference Concerning Several Redundancy, *Journal of Multivariate Analysis,* vol. 79, 71-88, 2001.

[LAZ 02] LAZRAQ, A., CLEROUX R.. Inférence Robuste sur un Indice de Redondance, *Revue de Statistique Appliquée*, vol. 4, 39-54, 2002.

[LEB 87] LEBART L., Conditions de Vie et Aspirations des Français, Evolution et Structure des Opinions de 1978 à 1984, *Futuribles*, vol.1, 25-26, 1987.

[LEB 97] LEBART, L., MORINEAU, A., PIRON, M.*, Statistique Exploratoire Multidimentionnelle 2^{e} edition,* DUNOD, 1997.

[LEB 91] LEBBE, J., *Représentations des Concepts en Biologie et en Médecine*, Thèse de PhD, Université Paris 6, avril 1991.

[LER 73] LERMAN, I.C, Etude Distributionnelle de Statistiques de Proximité entre Structures Finies de Mêmes Types; Application à la Classification Automatique, *Cahier no.19 du Bureau Universitaire de Recherche Opérationelle,* Institut de Statistique des Universités de Paris, 1973.

[LER 81] LERMAN, I.C., *Classification et Analyse Ordinale des Données*, Dunod, Paris, 1981.

[LER 88] LERMAN, I.C, Comparing Partitions (Mathematical and Statistical Aspects), *Classification and Related Methods of Data Analysis,* H.H Bock Editor, 121-131, 1988.

[MAH 30] MAHALANOBIS, P.C., On Tests and Measures of Groups Divergence I. *Journal of the Asiatic Society of Bengal,* vol. 26, 541, 1930.

[MAL 01] MALERBA, D., ESPOSITO, F., GIOVIALE, V., TAMMA, V. Comparing Dissimilarity Meausures for Symbolic data analysis. http://www.di.uniba.it/~malerba/publications/ntts-asso.pdf, 2001.

[MAR 84] MARCOTORCHINO F., Utilisation des Comparaisons par Paires en Statistique des Contingences (Partie II), *Etude du Centre Scientifique IBM France*, No F069, 1984.

[MAR 91] MARCOTORCHINO J.F., EL AYOUBI, N., Paradigme Logique Des Ecritures Relationnelles De Quelques Critères Fondamentaux D'Association*, Revue de Statistique Appliquée*, vol. 2, 25-46, 1991.

[MAT 55] MATUSIKA, K. On the Theory of Statistical Decision Functions. *Ann. Math. Stat*. vol. 26, 631-640, 1955.

[McC 87] MCCUTCHEON, A.L, *Latent Class Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 64 , SAGE publications, 1987.

[McI 80] MCINTYRE, R.M., BLASHFIELD R.K., A Nearest-Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure, *Multivariate Behavior Research*, vol. 15, 225-238, 1980.

[McL 97] MCLACHLAN, G.J., et KRISHNAN, J., *The EM algorithm and Extensions*, Wiley, New York, 1997.

[McL 00] MCLACHLAN, G.J., PEEL, D., *Finite Mixture Models*, Wiley, New York, 2000.

[MEH 03] MEDHI, L., DIDAY, E., WINSBERG, S., Symbolic Class Description with Interval Data, *the Electronic journal of Symbolic Data Analysis*, vol.1, No 1, 2003.

[MIC 81] MICHALSKI, R.S, DIDAY, E., Stepp, R.E.,- A Recent advance in data analysis: Clustering objects into classes characterized by conjunction concepts-Progress in *Pattern Recognition*, vol 1,North Holland, Amsterdam, pp 33-56, 1981.

[MIC 83] MICHALSKI, R.S., et STEPP, R.E, *Learning from Observations: Conceptual Clustering*, chap.4, vol. 1, 1983.

[MIL 85] MILLIGAN G.W., COOPER M.C., An Examination of Procedures for Determining the Number of Clusters in a Data set. *Psychometrika,* vol. 50, 159-179, 1985.

[MIL 86] MILLIGAN G.W., COOPER M.C., A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis, *Multivariate Behavior research*, vol. 21, 441-458, 1986.

[MIL 96] MILLIGAN G.W., CHENG, R., Measuring the Influence of Individual Data Points in a Cluster Analysis, *Journal of Classification*, vol. 46, 187-189, 1996.

[MOR 84a] MOREY, L.C., AGRESTI, A., The Measurement of Classification Agreement: An Adjustment of the Rand Statistic for Chance Agreement, Educational and Psychological Measurement, 44, 33-37, 1984.

[MOR 95] MORINEAU, A., SUMMA, M., TONG, H., Marquage Sémantique des Classes et des Axes- *28 èmes journées de l'ASU*, Paris, 468-472, 1995.

[MOR 84] MORINEAU, A., Note sur la Caractérisation Statistique d'une Classe par les Valeurs Tests, *Bulletin Technique de CESIA*, vol. 2 (1), 20-27,1984.

[NAK 00] NAKACHE, J.P., CONFAIS, J., *Méthodes de Classification avec illustration SPAD et SAS*. CISIA CERESTA, Montreuil, 2000.

[NAP 92] NAPOLI, A., *Représentations à Objets et Raisonnement par Classification en Intelligence Artificielle*, Nancy, Thèse de PhD, Université de Nancy 1, 1992.

[PER 96] PERINEL, E., *Segmentation en Analyse des Données Symboliques*, Thèse de PhD, Université Paris IX Dauphine, septembre 1996.

[POP 83] POPPING, R. Traces of agreement. On the Dot- Product As a Coefficient of Agreement. *Quality and Quantity*, vol. 17, No1, 1-18, 1983 .

[POP 84] POPPING, R. Traces of Agreement: On Some Agreement Measures for Open- Ended Questions, *Quality and Quantity*, vol. 18, No 2, 147-58, 1984.

[POP 88] POPPING, R. On Agreement Indices for Nominal Data. *in Sociometric research,, Edited by W.E Saris and I.N. Gallhofer, McMillan, London* vol. 1, 90-105, 1988.

[POP 92] POPPING, R. Taxonomy on Nominal Scale Agreement, *Groningen ; iec ProGAMMA*, 1945-1990, 1992.

[POP 00] POPPING, R. The Computer Program AGREE 7 for nominal Scale Agreement. http://www.ppsw.rug.nl/~popping/RP_131.html, 2000.

[RAN 71] RAND, W.M., Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, vol. 66 (336), 846-850, 1971.

[RAS 94] RASSON, J.P., KUBUSHISHI, T., The Gap test: An Optimal Method for Determining the Number of Natural Classes in *Cluster Analysis, New Approaches in Classification and Data Analysis*, Diday, E., et al., Eds., Springer Verlag, Berlin, 186-193, 1994.

[RIS 89] RISSANEN, J., *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Company, Teaneck, New Jersey, 1989.

[RIS 94] RISSON, A., ROLLAND, P., CHAUCHAT J.H, *Analyse Graphique d'une Matrice De Données*. Guide Pratique, CISIA, 1994.

[ROB 76] ROBERT P., ESCOUFIER, Y. A Unifying Tool for Linear Multivariate Statistical Methods: the RV-coefficient. *Appl. Statist*., vol. 25**,** 257-265, 1976.

[SAP 90] SAPORTA, G., *Probabilités Analyse des Données et Statistique*, Editions TECHNIP, 1990.

[SAP 97] SAPORTA, G. Problèmes Posés par la Comparaison de Classifications Dans des Enquêtes Différentes, *53 ème session de l'Institut International de Statistique,* Istanbul, août 1997.

[SAP 01] SAPORTA G., YOUNESS G. Concordance entre Deux Partitions: Quelques Propositions et Expériences, *in Proceedings SFC 2001, 8èmes rencontres de la Société Francophone de Classification,* Pointe à Pitre, 2001.

[SAP 02] SAPORTA G., YOUNESS G. Comparing Two Partitions: Some Proposals and Experiments, *Proceedings in Computational Statistics edited by Wolfgang Härdle, Physica- Verlag*, Berlin, Germany, 2002.

[SAS 94] SAS/ STAT, User's guide, version 6, fourth edition. SAS Institute Inc., Cary, NC (USA), 1994.

[SIL 02] SILVA, A.L, BACELAR- NICOLAU, H., SAPORTA, G., Missing Data in Hierarchical Classification of Variables- A Simulation Study, *Classification Clustering and Data analysis, Springer*, 121-128, 2002.

[SMI 80] SMITH, S.P., JAIN A.K., Testing for Uniformity in Multidimensional Data, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI 6, 73-91, 1980.

[SOK 58] SOKAL, R.R., MICHENER, C.D., A Statistical Method for Evaluating Systematic Relationships, *Univ Kansas Sci. Bull.,* vol. 38, 1409- 1438, 1958.

[SOK 63] SOKAL, R.R., SNEATH P.H.A. *Principles of Numerical Taxonomy,* Freeman and co., San Francisco, 1963.

[SOK 88] SOKAL, R.R., Unsolved Problems in Numerical Taxonomy in *Classification and Related Methods of Data Analysis*, *H.H Bock ed., North Holland*, 45-56, 1988.

[SOR 48] SORENSEN, T., A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarities of Species Content and its Application to Analyses of the Vegetation on Danish Commons. *Biologiske Skrifter*, vol.5, 1-34, 1948.

[SOU 02] SOUSA, A., SILVA, O., BACELAR-NICOLAU, H., NICOLAU, F., *Validação em Classificação Hierárquica. JOCLAD, 2002*.

[SPL 00] *S-PLUS 2000 User's Guide. Data Analysis Products Division*, MathSoft, Seattle, Washington. Ed. Springer, 2000.

[STE 98] STEPHAN, V., *Construction d'Objets Symboliques par Synthèse des Résultats de Requêtes SQL*, Thèse de PhD, Université Paris IX -Dauphine, 1998.

[STE 68] STEWART D., LOVE W., A General Canonical Correlation index, *Psychological Bulletin,* vol. 70, 160- 163, 1968.

[TOM 88] TOMASSONE, R., DANZART, M., DAUDIN, J.J., MASSON, J.P., *Discrimination et Classement*, Masson, Paris, 1988.

[VAN 71] VAN EMDEN, M.H*., An Analysis of Complexity*, Mathematical Center Tracts, vol. 35, Amsterdam, 1971.

[VEN 99] VENABLES W.N., RIPLEY B.D., S Programming, Statistics and Computing, *Ed. Springer*, 1999.

[VER 99] VERMUNT, J.K., MAGIDSON, J., Exploratory Latent Class Cluster, Factor and Regression Analysis :The latent Gold Approach . *Article présenté à la conférence de EMPS'99,*Lueneburg, Germany, 1999.

[VER 00] VERMUNT, J.K., MAGIDSON, J., Latent GOLD 2.0 User's Guide, Belmont, MA: Statistical Innovations Inc.

[VER 02] VERMUNT, J.K., MAGIDSON, J., *Latent Class Cluster Analysis*. In J.A. Hagenaars et A.L. McCutcheon Eds., Applied Latent Class analysis, 89- 106, Cambridge University Press., 2002.

[VIG 03] VIGNEAU, E., QUANNARI, E.M., Clustering of Variables Around Latent Components, *Communications in Statistics Simulation and Computation*, vol. 32 (4), 1131-1150, 2003.

[WON 85] WONG, M.A., A Bootstrap Testing Procedure for Investigating the Number of Subpopulations, *Journal Statistics Comput. And Simulations,* vol. 22, 99-112, 1985.

[YOU 03] YOUNESS G., SAPORTA G., Sur les Indices de Comparaison de Deux Classifications, *in Proceedings SFC 2003, 10èmes rencontres de la Société Francophone de Classification,* Neuchâtel, 2003.

[YOU 04] YOUNESS G., SAPORTA G., Une Méthodologie pour la Comparaison de Partitions, *Revue de Statistique Appliquée*, vol. LII (1), 97-120, 2004.

[YOU 04'] YOUNESS G., SAPORTA G., Some Measures of Agreement Between Close Partitions, Student, à paraître.