Université Paris 1 Panthéon-Sorbonne

THÈSE

Présentée par Smaïl IBBOU

Pour obtenir le grade de

Docteur en Sciences
(Arrêté du 30 mars 1992)

Spécialité : Mathématiques Appliquées

Titre: **Classification, analyse des correspondances et méthodes neuronales**

Sous la direction de Marie COTTRELL

Soutenue publiquement le 20 janvier 1998 devant le jury composé de

Mesdames	Marie COTTRELL	
	Anne GUÉRIN-DUGUÉ	Rapporteur
Messieurs	Eric de BODT	
	François GARDES	
	Jean-Pierre FÉNELON	
	Xavier GUYON	Président
	Michel VERLEYSEN	Rapporteur

Remerciements

Je voudrais tout d'abord remercier Marie Cottrell pour m'avoir offert la possibilité de traiter ce sujet et de m'avoir ensuite si bien encadré avec beaucoup de sérieux et de patience. Je peux dire sans exagérer que cette thèse est le fruit de son soutien permanent.

Je remercie également Anne Guérin-Dugué et Michel Verleysen d'avoir accepté d'être les rapporteurs de cette thèse. Je les remercie pour leur lecture profonde de ce document et pour leurs remarques constructives.

Je remercie particulièrement Xavier Guyon d'avoir accepté de présider mon jury.

J'adresse aussi tous mes remerciements aux autres membres du jury, à Eric de Bodt, Jean-Pierre Fénelon et François Gardes.

Au début de ce travail, j'ai eu le plaisir de collaborer avec Patrice Gaubert et Christian Tutin à la réalisation de l'étude DREIF. Je tiens à leur exprimer ma gratitude.

Un merci particulier à tous les membres du SAMOS qui m'ont offert un cadre de travail chaleureux et motivant.

Finalement j'adresse un salut fraternel à tous ceux qui m'ont entouré durant ce parcours et avec qui j'ai partagé beaucoup de moments agréables.

À mes parents, mes frères et mes sœures.

Table des matières

Chapitre 1

Introduction

Les réseaux de neurones ont d'abord été développés pour résoudre des problèmes de contrôle, de reconnaissance de formes ou de mots, de décision, de mémorisation comme une alternative à l'intelligence artificielle, et en relation plus ou moins étroite avec la modélisation de processus cognitifs réels et des réseaux de neurones biologiques. Dans le même temps, de nombreux auteurs ont remarqué qu'il existait des relations étroites entre les modèles neuronaux et la statistique (voir par exemple dans (Hertz et al., 1991), (Baldi and Hornik, 1989) et (Bourlard and Kamp, 1988)).

Ces relations sont de diverse nature. Dans certains cas, les modèles neuronaux réalisent des taches traditionnelles de la statistique, comme la classification (perceptron, modèle compétitif), la régression et la prévision (perceptron multicouches). D'autre part, les chercheurs ont proposé des méthodes connexionistes qui permettent de réaliser par des méthodes itératives des analyses de données classiques telles que l'analyse en composantes principales ((Oja, 1982), (Oja, 1989), (Sanger, 1989) et (Földiák, 1989)) et l'analyse discriminante ((Gallinari et al., 1988)).

Dans ce cadre, notre travail s'intéresse plus particulièrement à l'algorithme de Kohonen qui est un des algorithmes neuronaux les plus connus et utilisés, mais souvent de manière peu motivée et sans qu'on perçoive pourquoi il apporte plus qu'un simple algorithme de classification.

En fait les thèmes que nous avons abordés durant ce travail de thèse découlent directement des préoccupations rencontrées lors de notre stage de DEA. Celui-ci devait être à l'origine une application presque directe, sur des données socio-économiques recensées en Ile de France, de méthodes statistiques simples telles que celles de l'analyse des données. Cependant il se trouve que ces méthodes, qui sont réputées en théorie être des méthodes simples et efficaces, sont parfois en pratique insuffisantes ou inadéquates. Cette insuffisance peut être liée d'une part à la méthode elle même, telle que la linéarité des méthodes dites factorielles, et d'autre part à la base de données qui peut contenir des observations incomplètes auquel cas les méthodes classiques sont inutilisables. Nous avons donc pensé aux méthodes neuronales et en particulier à l'algorithme de Kohonen pour résoudre les problèmes posés.

Ce stage a consisté en une étude empirique pour le compte de la Direction Régionale de l'Equipement d'Ile de France sur le marché immobilier et foncier francilien. Il s'agissait de trouver des modèles pouvant mettre en évidence des relations éventuelles entre le prix du sol et les prix des logements dans les huit départements de la Région Parisienne. D'un point de vue purement économiste, il est assez intuitif de penser qu'il n'est pas possible de trouver un modèle unique de prix pour toute l'Ile de France du fait que le marché foncier et immobilier n'est pas homogène. Par exemple, en considérant deux cas extrêmes, la commune de Saint-Denis ne possède pas réellement de marché au sens spéculatif (le marché est plutôt administré), alors que la situation est bien différente dans des arrondissements de Paris tel que le huitième ou le seizième. D'ailleurs les premières tentatives d'élaboration de modèles linéaires sur l'ensemble des communes n'ont pas donné de résultats satisfaisants.

Cette hétérogénéité des situations nous a donc amené à faire d'abord une segmentation pour pouvoir ensuite trouver des modèles de prix plus robustes et plus significatifs. La littérature en économie urbaine suggère de faire une partition en trois segments identifiés par la suite comme «Supérieur», «Moyen» et «Banal». Mais cette tripartition n'est pas entièrement élucidée. En effet, si on prend comme exemple le segment Moyen, celui-ci est constitué d'une part d'éléments qui n'ont pas suffisamment de côte pour rejoindre le segment supérieur et d'autre part d'éléments qui ne sont pas totalement en bas de l'échelle. On peut donc supposer que ce segment est lui même réunion d'un segment « Moyen supérieur » et un autre « Moyen inférieur». Ce genre de problématique nous a conduit à étudier différentes méthodes de classification et de regroupement de classes. Nous avons étudié à ce sujet un algorithme de fusion de données dû à Wong Y.F. et paru dans (Wong, 1993), qui donne un critère assez naturel pour le choix du nombre pertinent de classes.

Un autre problème majeur était cependant inhérent à la base de données, car elle contenait une proportion importante de valeurs manquantes et donc rendait inefficaces les méthodes classiques d'analyse de données comme l'ACP ou la classification hiérarchique. Par exemple pour la segmentation, nous avons effectué sur la base de données, une ACP et une classification hiérarchique au moyen du logiciel SAS. Mais chaque fois que le programme rencontre une valeur manquante pour une ligne du tableau de données (qui correspond à une commune), toute la ligne est supprimée. Finalement en considérant 6 variables, il ne restait qu'une trentaine de communes sur les 256 recensées. En rajoutant des variables, on augmente le nombre des valeurs manquantes et donc on diminue encore le nombre de communes classées. Nous avons alors été conduit à adapter l'algorithme d'auto-organisation de Kohonen pour pouvoir faire la segmentation en tenant compte des valeurs manquantes. Cette méthode nous a permis de classer 200 communes avec 23 variables. Après cette expérience réussie, on s'est donc proposé de faire une étude empirique permettant de quantifier le degré de robustesse de l'algorithme de Kohonen face à des observations incomplètes. L'étude est menée sur des exemples constitués par des observations simulées et des observations issues d'une base de données réelles. On propose également par cette méthode un moyen d'estimer les valeurs manquantes.

L'autre aspect auquel on s'est intéressé est le traitement des variables qualitatives (ou discrètes) par l'algorithme de Kohonen. Les cartes de Kohonen semblent être tout à fait adaptées à ce type de traitement et donnent de très bons résultats. Le cas où l'on croise deux variables qualitatives dans une table de contingence a été traité dans (Cottrell et al.,

1993) par un algorithme appelé Kouplet. Nous avons proposé dans (Ibbou and Cottrell, 1995b) un méthode originale basée sur l'algorithme de Kohonen qui permet de classer, dans une carte de Kohonen, les modalités de 2 variables ou plus. L'étude a été poursuivie pour proposer finalement deux façons de classer des individus et des modalités dans une carte de Kohonen. La carte obtenue fait ressortir non seulement les groupements d'individus qui se ressemblent, mais elle associe également les modalités qui caractérisent chaque groupe d'individus. Le premier procédé utilise le tableau disjonctif complet et le second utilise la matrice de Burt. Cette deuxième méthode est de plus très avantageuse du point de vue du temps de calcul.

Le chapitre 2 de ce document traite de la classification et de l'étude de l'algorithme de fusion. Dans le chapitre 3 on rappelle les méthodes classiques de l'analyse en composantes principales et de l'analyse des correspondances. On expose ensuite dans le chapitre 4 l'algorithme de Kohonen sur des exemples simulés et des exemples réels. Le chapitre 5 présente l'algorithme de Kohonen adapté à la présence de valeurs manquantes et l'étude empirique faite sur cette méthode. Enfin nous présentons dans le chapitre 6 les méthodes **KACM I** et **KACM II** pour le traitement des variables qualitatives.

Chapitre 2

Classification et algorithme de fusion

2.1 Introduction

La classification des données est un domaine de la statistique appliquée qui intervient dans beaucoup de domaines et de disciplines tels que l'économie, la finance, les télécommunications (transmission de données, reconnaissance des formes), la médecine (l'imagerie, diagnostics et systèmes experts), la documentation (rangements de titres par mots clefs et recherche par système de requêtes) et bien d'autres applications. Le mot « *classification* » désigne généralement, dans l'immense littérature qui traite des diverses techniques, un processus permettant de regrouper des données par groupes et de constituer un ensemble de classes qui ont un « *sens* » et un « *intérêt particulier* ». Dans le mot « *sens* » on exprime l'idée suivante : un élément classé dans une classe A est plus « *proche* » (nous verrons plus loin le sens qu'on donne à ce terme et comment on peut le quantifier) des autres éléments de la classe A que des éléments appartenant à une autre classe B. D'autre part une classification est envisagée par rapport à des objectifs préalables qui vont justifier le choix de la méthode de classification et par conséquent influencer les résultats qui seront obtenus. C'est ce que l'on entend par l'expression « *intérêt particulier* ». Ainsi la qualité d'une méthode de classification peut être finalement mesurée uniquement par ses capacités à répondre au mieux aux objectifs fixés au départ.

En économétrie, souvent le but est de faire une segmentation de l'ensemble des individus (catégories d'entreprises, villes, ménages, portefeuilles, clients...) en groupes homogènes, qui nous permette d'établir des modèles explicatifs et prévisionnels robustes. En effet la structure micro-économique de certains marchés, la spécificité des entreprises des villes (Voir l'étude « *Marché foncier et immobilier en Île de France* ») rendent la construction de modèles globaux assez complexe dès que l'on souhaite faire intervenir dans les équations plusieurs variables exogènes qui, à priori, sont significatives dans des classes d'individus distinctes. Le moyen naturel de mesurer la qualité d'une classification est donc lié dans ce cas, à la qualité des modèles obtenus.

Dans le cas de la transmission des données et des télécommunications, (images, sons ...) l'objectif est d'optimiser la vitesse et le coût de l'acheminement des données. Celles ci sont subdivisées en classes codées chacune par un individu moyen déterminé par la méthode de classification (dans ce cas on parle plutôt de quantification vectorielle). Cet individu moyen sera seul transmis en tant que représentant de toute sa classe. On pourra alors comparer deux méthodes de quantification par l'économie réalisée en coût de transmission et par la qualité de la reconstitution des données originelles à l'arrivée.

Dans le contexte de la recherche de documents, qui signifie la recherche d'un titre ou d'un ensemble de titres à partir d'une requête, on n'envisage pas de tester la requête sur tous les éléments de la base de données car le temps d'exploitation croît considérablement avec le nombre de requêtes et de titres. Une classification des titres est indispensable afin de limiter la recherche, en premier lieu, uniquement aux représentants des différentes classes. La recherche est ensuite achevée sur les individus appartenant aux classes des représentants ayant obtenu un score élevé durant le test avec la requête. Là aussi la qualité de la classification repose sur les performances obtenues durant la recherche de documents.

Le but est donc de regrouper les données définies par un ensemble de variables dans des classes homogènes et distinctes entre elles. L'homogénéité d'une classe traduit le fait d'une part que les individus présents, a priori, dans une même classe possèdent des caractéristiques suffisamment semblables pour être considérés comme un seul individu et confondus avec leur représentant et d'autre part que deux individus, a priori, appartenant à deux classes différentes sont suffisamment distincts pour être affectés à deux représentants différents. Cela se traduit par une dispersion minimale des individus autour du centre de la classe. D'un autre côté, on désire que les distances entres les centres de classes soient maximales. En termes statistiques, cela équivaut à une variance intra-classes des variables aléatoires minimale et une variance interclasses maximale.

Il y a, à ce stade, une distinction à noter entre une classification et une quantification vectorielle qui tout de même restent intimement liées. Mais le fait de vouloir distinguer ces deux méthodologies résulte principalement du fait que certains algorithmes sont plutôt destinés à faire de la classification (comme la méthode des k-means) et d'autres plutôt destinés à faire de la quantification (comme le Learning Vector Quantisation). La première méthode a pour but d'abord de regrouper les données en classes homogènes dont un représentant peut être calculé ensuite, tandis que la seconde affecte chaque donnée à un représentant et la partition est automatiquement générée.

2.2 Définitions et notations

Les individus que nous voulons classer sont représentés par des vecteurs appartenant à $\mathbb{R}^d$. L'ensemble de ces individus est noté I de cardinal n et l'ensemble des représentants est noté $\hat{I}$ de cardinal m.

Définition 2.2.1
Un quantificateur d-dimensionnel d'ordre m est une application q d'un ensemble $I \subset \mathbb{R}^d$ d'individus, de cardinal n, vers un ensemble $\hat{I} \subset \mathbb{R}^d$ de *représentants*, de cardinal m avec $m << n$, qui à tout individu $x_i \in I$, $1 \leq i \leq n$ associe un représentant $y_j \in \hat{I}$, $1 \leq j \leq m$.

$$\begin{array}{ccc} I & \xrightarrow{q} & \hat{I} \\ x_i & \longrightarrow & q(x_i) = y_j \end{array}$$

- $\hat{I}$: est appelé le dictionnaire ou «codebook»
- $C_j = \{x_i,\ q(x_i) = y_j\}$ est la j-ème classe
- $\mathcal{C} = \{C_j,\ j = 1, ..., m\}$ est la partition en classes générée par q

Pour quantifier la proximité de deux individus, on introduit la notion de *distorsion* dans le cas où les variables décrivant la base de données sont des variables continues et la notion de « *similarité* » dans le cas où les variables sont discrètes ou qualitatives. Dans les deux cas, les mesures de proximité sont nombreuses dans la littérature et souvent dépendantes de la nature des données que l'on traite.

Dans le cas continu, la distorsion la plus fréquemment utilisée, surtout en pratique pour des raisons de vitesse de convergence, est celle relative à la norme euclidenne, appelée *erreur quadratique* ou *distorsion quadratique*.

$$d(x, y) = \|x - y\|_2 = \left(\sum_{k=0}^{k=d} (x[k] - y[k])^2 \right)^{\frac{1}{2}}$$

D'autres mesures de distorsion sont aussi utilisées selon les cas, en voici quelques unes :

erreur de la norme de Holder l_ν, $\nu \in \mathbb{N}^*$

$$d(x, y) = \|x - y\|_\nu = \left(\sum_{k=0}^{k=d} (x[k] - y[k])^\nu \right)^{\frac{1}{\nu}}$$

erreur de la norme infinie l_∞

$$d(x, y) = \|x - y\|_\infty = \max_{1 \leq k \leq d} (|x[k] - y[k]|)$$

erreur quadratique pondérée

$$d(x, y) = \left(\sum_{k=0}^{k=d} w_k (x[k] - y[k])^2 \right)^{\frac{1}{2}}$$

où les w_k sont des constantes de pondération

erreur quadratique généralisée

$$d(x, y) = (x - y)^t B(x - y)$$

avec B une matrice $k \times k$ symétrique définie positive.

Toutes ces distorsions dont nous avons rappelé brièvement les définitions ne dépendent que de la différence $(x - y)$. On les appelle «mesures de distorsion de la différence». Il existe d'autres mesures de distorsions qui sont des fonctions plus complexes de x et de y (i.e pas uniquement de $(x - y)$).

$$d(x, y) = (x - y)^t R(x)(x - y)$$

où $R(x)$ est une matrice $k \times k$ symétrique définie positive, dépendant de x.

A partir de ces mesures de distorsion, on définit une mesure de qualité du quantificateur et par conséquent un critère de comparaison de deux quantificateurs.

2.2.1 Performance

Dans la suite, on ne considère que des données quantitatives et l'erreur quadratique. On cherche à définir un critère quantitatif permettant d'évaluer la qualité d'un quantificateur vectoriel et par conséquent un outil de comparaison entre deux quantifications vectorielles. Munissons l'ensemble des données d'un modèle probabiliste et notons $X = (X^1, ..., X^d)$ un vecteur aléatoire dont les réalisations sont les vecteurs $x_i = (x_i^1, ..., x_i^d)$, $1 \leq i \leq n$ de l'ensemble I. La loi du vecteur aléatoire est définie par sa fonction de répartition

$$F(x) = P(X^i \leq x^i,\ i = 1, ..., d), \quad x = (x^1, ..., x^d) \in \mathbb{R}^d$$

Alors une mesure D de la performance du quantificateur q peut être définie par l'espérance de la distorsion entre X et $q(X)$, pour la loi de probabilité P de X.

$$D(q) = E[d(X, q(X))] = \int d(X, q(X))dP(X) \tag{2.2.1}$$

Si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes de même loi que X, alors d'après la loi des grands nombres

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i, q(X_i)) \xrightarrow[n\to\infty]{ps} D(q)$$

et si $(x_i,\ 1 \leq i \leq n)$ sont n observations indépendantes de X, on peut estimer $D(q)$ par

$$\frac{1}{n}\sum_{i=1}^{n} d(x_i, q(x_i))$$

Une autre mesure de performance qu'on peut utiliser est le maximum calculé sur tout l'ensemble I des distorsions de chaque élément de cet ensemble à son centre correspondant.

$$D(q) = \max_{x_i \in I} d(x_i, q(x_i))$$

L'objectif d'une bonne quantification est évidemment de minimiser la fonction de coût pour avoir la meilleure correspondance possible entre les vecteurs d'entrée et leur représentants. Le problème donc est de trouver de bons algorithmes qui arrivent à atteindre les minima globaux si possible ou à défaut les meilleurs minima locaux.

Définition 2.2.2 (Quantificateur optimal)
Un quantificateur vectoriel q^* d'ordre m sera dit optimal (ou globalement optimal) s'il minimise l'espérance de la distorsion définie dans 2.2.1, c'est-à-dire :

$$D(q^*) \leq D(q) \quad \forall q, \text{ quantificateur d'ordre } m.$$

Plusieurs algorithmes ont été proposés dans la littérature, en vue de construire des quantificateurs vectoriels, pour diverses distributions de probabilité classiques pour P (si celle ci est bien évidemment connue) et pour différentes mesures de distorsion. Dans le chapitre suivant nous donnerons à titre d'exemple deux algorithmes de classification.

Lloyd proposa en 1957 (Lloyd, 1957) deux méthodes pour la quantification vectorielle dans le cas unidimensionnel ($d = 1$) et pour la distorsion quadratique. La méthode [II] est une approche variationnelle qui consiste à dériver l'espérance de la distorsion par rapport aux représentants y_j en restant dans l'enveloppe convexe de l'ensemble des données. En annulant ensuite les dérivées, cette méthode conduit généralement à des points stationnaires qui vérifient les conditions nécessaires mais pas suffisantes d'optimalité. Enfin en étudiant la dérivée seconde, on peut montrer qu'un tel point stationnaire est finalement un minimum local. De plus Lloyd a montré, dans certains cas, l'optimalité globale en cherchant d'une façon exhaustive tous les minima locaux.

Par contre la méthode [I] que Lloyd a proposé dans (Lloyd, 1957) est une approche non variationnelle qui donne dans certains cas les mêmes résultats que ceux de la deuxième méthode et qui a été généralisée au cas multidimensionnel par (Linde et al., 1980).

2.3 Deux algorithmes multidimensionnels

2.3.1 Algorithme LBG de Linde, Buzo et Gray

C'est un algorithme de quantification vectorielle qui a été proposé en 1980 par Linde, Buzo et Gray (Linde et al., 1980). C'est la généralisation de l'algorithme de Lloyd au cas multidimensionnel.

On suppose que la distribution de probabilité de la variable aléatoire X est connue, elle peut être discrète ou continue, et sans aucune hypothèse de dérivabilité.

Etant donné un quantificateur q, un ensemble de représentants $\hat{I} = \{y_j,\ 1 \leq j \leq m\}$ et une partition $\mathcal{C}$ de l'ensemble I des individus, $\mathcal{C} = \{C_j,\ 1 \leq j \leq m\}$, l'espérance de la distorsion s'écrit :

$$\begin{aligned} D(\hat{I},\mathcal{C}) &= E[d(X,q(X))] \\ &= \sum_{j=1}^{j=m} \int_{C_j} d(X,y_j)dP(X) \\ &= \sum_{j=1}^{j=m} E[d(X,y_j)|X \in C_j]P(X \in C_j) \end{aligned}$$

Supposons qu'on se donne un ensemble de représentants $\hat{I}$ et on cherche alors à trouver la partition $\mathcal{C}$ correspondant à cet ensemble $\hat{I}$ et qui minimise l'espérance de la distorsion D. La partition optimale pour cet ensemble $\hat{I}$ est celle qui affecte chaque $x_i \in I$ au représentant y_j, le plus proche au sens de $d(x_i, y_j)$.

$$x_i \xrightarrow{q} y_j = Arg \min_{y_l \in \hat{I}} d(x_i, y_l)$$

C'est la méthode du *plus proche voisin*. La partition correspondante s'appelle partition de Voronoï. Une condition d'arrêt pour un algorithme qui chercherait à calculer cette partition serait le fait qu'un élément de l'ensemble I doit être affecté à deux classes distinctes.

Soit la partition

$$P(\hat{I}) = \{P_j,\ j = 1, ..., m\}$$

avec P_j définie de la manière suivante,

$$x \in P_j \text{ (ou encore } q(x) = y_i) \text{ si et seulement si } d(x,y_j) \leq d(x,y_l)\ \forall l,\ 1 \leq l \leq m.$$

Alors

$$D(\hat{I}, P(\hat{I})) = E[\min_{y_j \in \hat{I}} d(X,y_j)] \tag{2.3.2}$$

et donc pour une autre partition $\mathcal{C}$

$$D(\hat{I},\mathcal{C}) = E[d(X,q(X))] \geq E[\min_{y_j \in \hat{I}} d(X,y_j)] = D(\hat{I}, P(\hat{I})) \tag{2.3.3}$$

Une telle partition $P(\hat{I})$ est donc celle qui minimise la distorsion moyenne D.

Réciproquement, soit $\mathcal{C} = \{C_j, j = 1, ..., m\}$ une partition associée à un quantificateur q. On suppose que la mesure de distorsion et la mesure de probabilité sont telles que pour tout ensemble C de probabilité non nulle (C est un sous ensemble d'un espace euclidien de dimension d) il existe un vecteur $y(C)$ tel que

$$E[d(X, y(C)|X \in C)] = \min_{u \in \mathbb{R}^d} E(d(X, u)|X \in C). \tag{2.3.4}$$

Ce vecteur $y(C)$ de $\mathbb{R}^d$ est appelé *centre de classe*, parfois *centre de gravité* ou *barycentre* de la classe C. Il est optimal pour une partition $\mathcal{C}$ donnée. Si pour tous les éléments de $\mathcal{C}$ un tel vecteur existe, alors il est clair que pour cette partition $\mathcal{C}$, il n'existe pas de *codebook* $\hat{I}$, $\hat{I} = \{y_1, ..., y_m\}$ qui conduise à une distorsion moyenne plus faible que celle générée par le *codebook* $y(\mathcal{C})$ où :

$$y(\mathcal{C}) \overset{\text{déf}}{=} \{y(C_j);\ 1 \leq j \leq m\}. \tag{2.3.5}$$

En effet

$$\begin{aligned} D(\hat{I}, \mathcal{C}) &= \sum_{j=1}^{j=m} E[d(X, y_j)|X \in C_j]\ P(X \in C_j) \\ &\geq \sum_{j=1}^{j=m} \min_{u \in \mathbb{R}^d} E[d(X, u)|X \in C_j]\ P(X \in C_j) \\ &= D(y(\mathcal{C}), \mathcal{C}). \end{aligned}$$

Dans l'équation 2.3.4, on doit donc rechercher le *codebook* $y(\mathcal{C}) = (y(C_1), y(C_2), ..., y(C_n))$ qui minimise l'espérance de la distorsion. Cette recherche se fait grâce aux méthodes classiques de minimisation, telles que la méthode de descente du gradient. On peut écrire l'algorithme dans le cas où la distribution des points d'entrée est supposée connue et dans le cas où elle ne l'est pas, auquel cas on utilise la distorsion empirique.

On fixe le nombre de classes m et un réel ϵ positif et petit.

Algorithme (distribution connue)

(0) étape d'initialisation	On choisit un codebook initial $\hat{I}_0$ et on pose $t = 0$ et $D_{-1} = \infty$
(1) une étape de calcul	A l'instant t, on a $\hat{I}_t = \{y_i, i = 1, ..., m\}$ on cherche la partition $P(\hat{I}_t) = \{C_i, i = 1, ..., m\}$ telle que $x \in C_i$ si $d(x, y_i) \leq d(x, y_j)\ \forall j$ on calcule $D_t = D(\hat{I}_t, P(\hat{I}_t)) = E[\min_{y \in \hat{I}_t} d(X, y)]$
(2) condition d'arrêt	Si $(D_{t-1} - D_t)/D_t \leq \epsilon$ on s'arrête, sinon on continue
(3) étape finale	On cherche le codebook optimal $\hat{I}_{t+1} = y(P(\hat{I}_t))$ $t := t + 1$ retour à (1)

Algorithme (distribution inconnue)

L'algorithme pour une distribution inconnue est le même que celui dans le cas d'une distribution connue. En effet ce qui change c'est le calcul de la distorsion qu'on approxime par l'erreur empirique,

$$D_t = \frac{1}{n} \sum_{j=1}^{n} \min_{y \in \hat{I}_t} d(x_j, y),$$

qui n'est autre que la variance intra-classes. Ces deux algorithmes donnent de bons quantificateurs vectoriels, mais dépendent fortement de l'initialisation. C'est la version appelée « batch » de l'algorithme k-means présentée ci-dessous.

2.3.2 Algorithme k-means

C'est un algorithme séquentiel efficace pour de longues séries de points et qui peut être utilisé pour initialiser d'autres algorithmes. C'et la version dite « online » de l'algorithme précédent. Il a été proposé initialement par (MacQueen, 1967) en 1967.

Comme pour l'algorithme précédent,l'objectif dans cette méthode est de trouver une partition $\mathcal{C} = \{C_1, ..., C_m\}$ de l'ensemble de départ $I = \{x_i, 1 \leq i \leq n\}$. A chaque étape de calcul, le *codebook* est constitué des centres de gravité (minima pour la distorsion euclidienne) des éléments C_i de $\mathcal{C}$. La partition finale pourra servir alors d'initialisation d'autres algorithmes.

Initialement on place les m premiers vecteurs $x_1, ..., x_m$ de l'ensemble I dans les m parties $C_1, ..., C_m$ de $\mathcal{C}$. Ce qui équivaut à initialiser le *codebook* avec les m premiers vecteurs de la base de données. A chaque itération, on observe une donnée x_i qu'on classera dans la classe C_j dont le centre de gravité $y(C_j)$ est le plus proche de x_i. On met à jour le centre de gravité $y(C_j)$ et on réitère cette opération jusqu'à ce que l'on ait classé tous les vecteurs appartenant à I. A La fin de l'apprentissage, on obtient une assez bonne partition de l'ensemble I mais qui peut être très différente de celle produite par l'algorithme L.B.G.

En effet même si l' affectation des données observées à leurs classes correspondantes, ainsi que la mise à jour du codebook sont identiques dans les deux cas, il est clair qu'il y a une différence fondamentale qui réside dans le fait que le premier algorithme regarde à chaque itération toute la base de données par contre le second ne prend en considération que la donnée observée. Cependant si n tend vers l'infini les deux algorithmes sont équivalents. Ainsi, dans la méthode k-means, si n n'est pas grand alors, pour faire converger l'algorithme, on peut représenter plusieurs fois les individus. C'est la version stochastique de L.B.G.

2.3.3 La fusion des données

D'autres algorithmes sont disponibles et les uns plus utilisés que d'autres. Comme bonnes références, on peut citer les livres de (Duda and Hart, 1973), (Jain and Dubes, 1988) et (Fukunaga, 1990) où l'on peut trouver un panorama de méthodes telles que les méthodes bayesiennes sur des mélanges gaussiens ou le percepton multicouches. Mais tous ces algorithmes considèrent que le nombre de classes est fixé a priori. Cela représente souvent une limitation importante et arbitraire. On s'est donc particulièrement intéressé à un algorithme de classification basé sur le principe de fusion des données. Cet algorithme introduit par (Wong, 1993) utilise une méthode d'estimation par maximum d'entropie et donne un critère significatif pour le choix du nombre de classes. Notre but est de formaliser le problème et de tester l'algorithme en comparaison avec d'autres algorithmes sur des données réelles. Nous commençons d'abord par faire un bref rappel sur l'information de Kullback et sur l'entropie.

2.4 Information de Kullback. Entropie

2.4.1 Information de Kullback, (entropie relative)

On suppose que tout phénomène aléatoire est régi par une loi de probabilité inconnue P. La modélisation du phénomène consiste à estimer la loi de probabilité P dans un but de simulation ou de prédiction. Pour un minimum de contrôle, on est amené à mesurer l'écart qu'il y a entre la probabilité estimée P et une loi de probabilité de référence Q. Souvent cette loi de probabilité Q n'est autre que la loi empirique observée du phénomène. Nous devons donc avoir un critère pour mesurer la dissemblance entre une loi P et la loi Q. On dispose de plusieurs critères tels que la distance en variation, la distance du Chi-deux, la distance de Hellinger ou l'information de Kullback. On s'intéresse particulièrement à ce dernier critère en vue de l'appliquer à la classification de données.

Soient $(\Omega, \mathcal{A}, \mu_i)$, $\quad i = 1, 2$ deux espaces de probabilité et on suppose que les probabilités μ_1 et μ_2 sont équivalentes, c'est-à-dire que μ_1 est absolument continue par rapport à μ_2 et réciproquement. Soit λ une probabilité dominant μ_1 et μ_2 (par exemple $\lambda = (\mu_1 + \mu_2)/2$).

On appelle f_i la densité de μ_i par rapport à λ. On a alors $d\mu_1 = f_1(x)d\lambda$ et $d\mu_2 = f_2(x)d\lambda$. Soit X une variable aléatoire sur Ω telle que sous l'hypothèse H_1, X suit la loi μ_1 et sous l'hypothèse H_2, suit la loi μ_2, alors d'après la formule de Bayes :

$$P(H_i|x) = \frac{P(H_i)f_i(x)}{P(H_1)f_1(x) + P(H_2)f_2(x)}, \qquad i = 1,2 \tag{2.4.6}$$

d'où l'on obtient

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1|x)}{P(H_2|x)} - \log \frac{P(H_1)}{P(H_2)} \tag{2.4.7}$$

Le membre de droite de l'égalité 2.4.7 représente la différence entre le log-ratio en faveur de H_1 après observation de x et le log-ratio en faveur de H_1 avant observation de x. Cette quantité peut être positive ou négative et elle peut être vue comme l'information qu'apporte l'observation de x. Le logarithme du rapport de vraisemblance, $\log[f_1(x)/f_2(x)]$ est *l'information en $X = x$ qui discrimine en faveur de H_1 contre H_2*, voir (Kullback, 1959).

On appelle *information de Kullback en faveur de H_1 contre H_2* et qui est notée $K(\mu_1, \mu_2)$ ou plus simplement $K(1,2)$ la quantité

$$K(\mu_1,\mu_2) = \mathbb{E}_{\mu_1}\left(\log \frac{f_1(X)}{f_2(X)}\right) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)$$

On définit plus généralement l'information de Kullback pour des lois de probabilités μ_1, μ_2 quelconques (qui peuvent être éventuellement étrangères) par :

$$K(\mu_1,\mu_2) = \begin{cases} \displaystyle\int \log \frac{\mu_1}{\mu_2}\, d\mu_1 & \text{si } \mu_1 \ll \mu_2 \\ 0 & \text{sinon} \end{cases}$$

avec les conventions suivantes :

$$0 \log \frac{0}{0} = 0 \quad 0 \log 0 = 0$$

Une propriété fondamentale est que $K(\mu_1, \mu_2) = 0$ si et seulement si $\mu_1 = \mu_2$, μ_1-presque sûrement. Voyons comment elle se traduit sur un exemple simple.

Exemple 2.4.1 Supposons que Ω est l'espace euclidien $\mathbb{R}^2$ et que $X = (X_1, X_2)$ un vecteur gaussien centré. On suppose que sous l'hypothèse H_1, les variables X_1 et X_2

sont dépendantes avec une densité jointe $f(x_1, x_2)$ et que sous l'hypothèse H_2, les deux composantes de X sont indépendantes. La matrice de variance-covariance de (X_1, X_2) est

$$\Gamma_{X_1,X_2} = \begin{bmatrix} (\sigma_{x_1})^2 & \varrho\sigma_{x_1}\sigma_{x_2} \\ \varrho\sigma_{x_1}\sigma_{x_2} & (\sigma_{x_2})^2 \end{bmatrix}$$

Sous l'hypothèse H_1, la loi de X a pour densité

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{(1-\varrho^2)}} \exp\left[-\frac{1}{2(1-\varrho^2)}\left(\frac{x_1^2}{\sigma_{x_1}^2} - 2\varrho\frac{x_1x_2}{\sigma_{x_1}\sigma_{x_2}} + \frac{x_2^2}{\sigma_{x_2}^2}\right)\right].$$

Sous l'hypothèse H_2, X a pour densité $f_{X_1}(x_1)f_{X_2}(x_2)$ où

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{x_1}} \exp\left(-\frac{x_1^2}{2\sigma_{x_1}^2}\right) \qquad f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{x_2}} \exp\left(-\frac{x_2^2}{2\sigma_{x_2}^2}\right).$$

Alors

$$K(1,2) = \int_{\mathbb{R}^2} f(x_1, x_2) \log \frac{f(x_1, x_2)}{f_{X_1}(x_1)f_{X_2}(x_2)} dx_1dx_2 = -\frac{1}{2}\log(1-\varrho^2).$$

On voit que $K(1,2)$ est nulle si $\varrho = 0$, c'est-à-dire quand les deux composantes sont indépendantes. A l'inverse lorsque les deux composantes sont de corrélation 1 ou -1, l'information $K(1,2)$ est infinie.

Exemple 2.4.2 Dans le cas discret, soit $\Xi = \{x_1, ..., x_n\}$ un ensemble fini, et soit $P = (p_1, ..., p_n)$ et $Q = (q_1, ..., q_n)$ deux mesures de probabilités sur Ξ. L'information de Kullback de P contre Q s'écrit

$$K(P,Q) = \sum_{i=1}^{n} \log \frac{p_i}{q_i} p_i$$

Cette quantité $K(P,Q)$ est nulle si et seulement si $\forall i, \ \ 1 \leq i \leq n, \ p_i = q_i$.

2.4.2 Entropie

On ne considère dans la suite que des variables discrètes finies, puisque l'on s'intéresse aux problèmes de classification d'un nombre fini d'individus. Soit X une variable aléatoire discrète à valeurs dans un ensemble $\Xi = \{x_1, ..., x_n\}$ de loi $P = (p_1, ..., p_n)$ avec $\sum_{i=1}^{n} p_i = 1$ et $0 \leq p_i \leq 1$. On définit l'entropie de la variable X ou de sa loi P par la quantité $S(X)$ ou $S(P)$ égale à :

$$S(P) = S(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{2.4.8}$$

avec la convention $0 \log 0 = 0$. On a $S(P) \geq 0$ et $S(P) = 0$ si et seulement si il existe un unique indice i tel que $p_i = 1$. D'autre part, il existe un lien entre l'information de Kullback et l'entropie de la loi P. En effet si U désigne la loi uniforme sur Ξ, alors

$$K(P, U) = \sum_{i=1}^{n} \log \frac{p_i}{1/n} p_i = \log n + \sum_{i=1}^{n} p_i \log p_i = \log n - S(P) \tag{2.4.9}$$

L'entropie de P est donc à la constante $+\log n$ près égale à l'opposé de la distance de Kullback entre P et la loi uniforme.

En théorie de l'information, cette quantité S, qu'on appelle mesure de Shannon, mesure le « taux d'incertitude » relatif à l'aléatoire uniforme.

Exemple 2.4.3 Cas d'une expérience de Bernoulli, $X \sim \mathcal{B}(p)$.

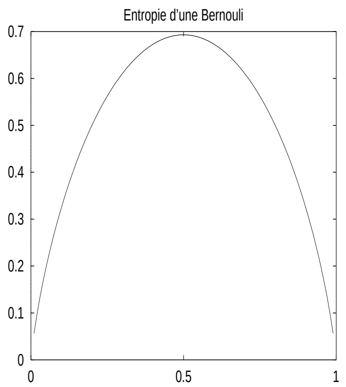
$$\text{On a} \quad P = (p_1, p_2) \text{ avec } p_1 = p \text{ et } p_2 = 1 - p,\ 0 \leq p \leq 1$$

$$S(p) = -p \log p - (1-p) \log(1-p)$$

$$S'(p) = -\log p + \log(1-p) \quad \text{et,}$$

$$S'(p) = 0 \Leftrightarrow p = \tfrac{1}{2}.$$

$$\text{De plus} \quad S''(p) = -1/p(1-p) < 0,\ \text{pour tout } p$$



Donc la fonction $S(p)$ est concave et atteint son maximum quand $p = \frac{1}{2}$, c'est-à-dire quand les événements « pile » et « face » ont les mêmes probabilités d'apparition. C'est bien d'une notion d'incertitude dont il s'agit.

Plus généralement si on cherche à optimiser $S(P)$ dans 2.4.8 sous la contrainte $\sum_{i=1}^n p_i = 1$, on montre aisément que $S(P)$ est une fonction concave dont le maximum est égal à $\log n$ et est atteint quand $p_i = \frac{1}{n}$ pour tout i, c'est-à-dire quand X suit une loi uniforme sur Ξ. En effet si on cherche à résoudre le problème ci-dessous

$$\begin{cases} \max\limits_{p_i} \left(-\sum\limits_{i=1}^{n} p_i \log p_i \right) \\ \sum\limits_{i=1}^{n} p_i = 1 \end{cases}$$

En utilisant la méthode des multiplicateurs de Lagrange, on considère le lagrangien

$$\mathcal{L}(p_1, ..., p_n, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda(\sum_{i=1}^{n} p_i - 1), \qquad \lambda \in \mathbb{R}$$

$$\tfrac{\partial \mathfrak{L}}{\partial p_i} = 0 \quad \Longrightarrow \quad p_i = e^{\lambda - 1}$$

donc tous les p_i sont égaux et donc égaux à $\dfrac{1}{n}$. De plus

$$\forall i \ \frac{\partial^2 \mathcal{L}}{\partial p_i^2} = -1 \qquad \text{et} \qquad \forall i, j \quad i \neq j \ \ \frac{\partial^2 \mathcal{L}}{\partial p_i \partial p_j} = 0$$

il s'agit donc bien d'un maximum qui vaut :

$$S(\frac{1}{n}, ..., \frac{1}{n}) = \log n$$

Remarquons également que le maximum de l'entropie $\log n$ croit avec la taille de Ξ. Finalement chercher à maximiser l'entropie sous l'unique contrainte de normalisation (ie $\sum p_i = 1$) équivaut à donner à tous les événements possibles la même probabilité. On retrouve le fait que $K(P, U) = 0$ si et seulement si P est la loi uniforme.

En réalité, on s'intéresse à la maximisation de l'entropie sous d'autres contraintes que celles qui sont liées à la normalisation, ce qui aboutit à des résultats beaucoup plus intéressants.

2.5 Estimation par maximum d'entropie

Supposons que $\Xi = \{x_1, ..., x_n\}$ soit le support d'une variable aléatoire discrète X. On note p_i la probabilité $P(X = x_i)$ et $P = (p_1, .., p_n)$ la loi de X. La loi P étant inconnue, on cherche à estimer les probabilités p_i. On suppose de plus qu'on a observé la moyenne d'une fonction $g(X)$

$$\mathbb{E}[g(X)] = \sum_{i=1}^{n} p_i g(x_i). \tag{2.5.10}$$

On part de l'équation 2.5.10 ci-dessus à laquelle on ajoute l'équation de normalisation

$$\sum_{i=1}^{n} p_i = 1 \tag{2.5.11}$$

Le problème qui consiste à estimer la loi P est à n inconnues, et nécessite donc pour sa résolution $(n-2)$ équations supplémentaires. Le problème peut, cependant, avoir une solution, en utilisant la maximisation de vraisemblance et les méthodes bayesiennes, si on pense avoir une idée sur un modèle a priori pour la loi P.

Mais on suppose que nous ne possédons pas d'information suffisante permettant de préférer un modèle par rapport à un autre, ce qui permettrait d'utiliser les méthodes d'estimation classiques. Cette idée de « *la raison insuffisante* », qu'on retrouve en théorie de l'information, est la principale motivation qui conduit à utiliser le concept de « *maximisation d'entropie* » pour modéliser cette incertitude. Ce principe de l'entropie maximale est vue, en théorie de l'information, comme l'estimateur le moins pénalisant au regard de l'information disponible (voir (Jaynes, 1957)).

La recherche des probabilités p_i devient alors un problème de maximisation sous contraintes qu'on peut résoudre en utilisant la méthode des multiplicateurs de Lagrange. On cherche à résoudre

$$\begin{cases} \sup\limits_{p_1,\ldots,p_n} \left(-\sum\limits_{i=1}^{n} p_i \log p_i \right) \\ \sum\limits_{i=1}^{n} p_i g(x_i) = C \ (C \text{ constante}) \\ \sum_{i=1}^{n} p_i = 1 \end{cases} \tag{2.5.12}$$

La méthode de l'entropie maximale est une méthode qui permet de reconstruire une densité de probabilité à partir d'information incomplètes, voir par exemple (Dacunha-

Castelle and Gamboa, 1990). On maximise $S(P)$ où P est une loi élément d'un ensemble $\mathcal{P}$ défini par des contraintes. Dans notre cas, c'est la contrainte 2.5.10 qui définit l'ensemble $\mathcal{P}$. Il s'agit d'une contrainte linéaire en P. La démonstration de l'existence et de l'unicité de la solution de ce problème a été traitée par Csiszar dans (Csiszar, 1970).

Dans notre cas la résolution du problème 2.5.12 n'est pas complexe. On utilise la méthode de Lagrange. Pour des raisons de calculs, destinées à simplifier l'écriture des formules, on utilise des multiplicateurs de Lagrange sous la forme λ et $(\mu - 1)$. L'expression du lagrangien $\mathcal{L}$ est alors :

$$\mathcal{L}(P, \mu, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda \left(C - \sum_{i=1}^{n} p_i g(x_i) \right) + (\mu - 1) \left(1 - \sum_{i=1}^{n} p_i \right) \tag{2.5.13}$$

et il s'agit de maximiser $\mathcal{L}(\mathcal{P}, \mu, \lambda)$.

$$\text{On a} \quad \forall i, \quad 1 \leq i \leq n, \quad \frac{\partial \mathcal{L}}{\partial p_i} = -\log p_i - \lambda g(x_i) - \mu$$

$$\text{et donc} \quad \forall i, \quad 1 \leq i \leq n, \quad \frac{\partial \mathcal{L}}{\partial p_i} = 0 \quad \Longrightarrow \quad p_i = e^{-\lambda g(x_i) - \mu}$$

$$\text{Or} \quad \sum_{i=1}^{n} p_i = 1 \quad \text{donc} \quad \forall i, \quad 1 \leq i \leq n, \quad p_i = \frac{e^{-\lambda g(x_i)}}{\sum_{i=1}^{n} e^{-\lambda g(x_i)}}$$

On appelle fonction de partition, la fonction

$$Z(\lambda) = \sum_{i=1}^{n} e^{-\lambda g(x_i)}.$$

On a alors,

$$\frac{\partial}{\partial \lambda} \log Z(\lambda) = -\frac{1}{Z(\lambda)} \sum_{i=1}^{n} e^{-\lambda g(x_i)} g(x_i) = -\sum_{i=1}^{n} p_i g(x_i) = -C$$

$$\text{soit} \quad C = -\frac{\partial}{\partial \lambda} \log Z(\lambda)$$

d'autre part

$$1 = \sum_{i=1}^{n} p_i = e^{-\mu} \left(\sum_{i=1}^{n} e^{-\lambda g(x_i)} \right) = e^{-\mu} Z(\lambda)$$

$$\text{soit} \quad \mu = \log Z(\lambda)$$

En résumé, on aboutit aux équations suivantes :

$$\left| \begin{array}{c} Z(\lambda) = \sum_{i=1}^{n} e^{-\lambda g(x_i)} \\ \\ \forall i, \quad 1 \leq i \leq n \quad p_i = \frac{e^{-\lambda g(x_i)}}{Z(\lambda)} \\ \\ C = -\frac{\partial}{\partial \lambda} \log Z(\lambda) \\ \\ \mu = \log Z(\lambda) \end{array} \right. \tag{2.5.14}$$

La solution du problème est donc fonction de la constante C correspondant à l'observation que l'on possède sur le système. La résolution complète nécessite généralement des méthodes numériques puisque la variable λ est une fonction implicite. En fait, au lieu de fixer C comme paramètre et d'obtenir λ implicitement, il est plus commode de prendre celui-ci comme paramètre. La forme des solutions p_i suggère alors de ramener le problème à la minimisation d'une énergie libre.

2.6 L'énergie libre en Mécanique Statistique

En physique statistique, on retrouve ces probabilités p_i formulées par 2.5.14 sous le nom de « *facteurs de Boltzmann* ». Ils sont obtenus en minimisant l'énergie libre d'un système physique évoluant dans n états possibles. Chaque état i est caractérisé par une probabilité d'existence p_i et une énergie E_i fonction de paramètres externes tels que le volume, la température, le potentiel, etc. L'énergie globale du système est la moyenne pondérée, par les probabilités des états, de toutes les énergies ; elle est définit par :

$$E = \sum_{i=1}^{i=n} p_i E_i$$

Si T est la température du système et $S = -\sum_{i=1}^{i=n} p_i \log p_i$ son entropie, on définie énergie libre E_l du système par :

$$E_l(p_1, ..., p_n) = E - TS$$

La minimisation de l'énergie libre E_l sous la contrainte de normalisation 2.5.11 ($\sum_{i=1}^n p_i = 1$) aboutit aux mêmes résultats que ceux obtenus par la méthode la maximisation sous contraintes de l'entropie. En effet, si l'on cherche à résoudre le problème :

$$\left\{ \begin{array}{l} E_l^* = \min\limits_{p_1,\ldots,p_n} \left(\sum\limits_{i=1}^n p_i E_i + T \sum\limits_{i=1}^n p_i \log p_i \right) \\ \\ \text{avec} \quad \sum_{i=1}^n p_i = 1. \end{array} \right. \tag{2.6.15}$$

L'expression du lagrangien $\mathcal{L}$ est :

$$\mathcal{L}(P,\mu) = \sum_{i=1}^n p_i E_i + T \sum_{i=1}^n p_i \log p_i + \mu \left(\sum_{i=1}^n p_i - 1 \right) \tag{2.6.16}$$

$$\forall i, \quad 1 \leq i \leq n, \quad \frac{\partial \mathcal{L}}{\partial p_i} = E_i + T \log p_i + T + \mu$$

$$\forall i, \quad 1 \leq i \leq n, \quad \frac{\partial \mathcal{L}}{\partial p_i} = 0 \quad \Longrightarrow \quad p_i = e^{-1-\mu/T} e^{-E_i/T}$$

$$\sum_{i=1}^n p_i = 1 \quad \Longrightarrow \quad \forall i, \quad 1 \leq i \leq n, \quad p_i = \frac{e^{-E_i/T}}{\sum_{j=1}^n e^{-E_i/T}}$$

On retrouve exactement les expressions attendues des probabilités p_i, (avec $T = \frac{1}{\lambda}$ et $E_i = g(x_i)$). La fonction de partition s'écrit alors $Z(T) = \sum_{i=1}^n e^{-E_i/T}$, et donc :

$$\forall i, \quad 1 \leq i \leq n, \quad p_i = \frac{e^{-E_i/T}}{Z}(T)$$

On remarque que le problème de maximisation de l'entropie sous une contrainte linéaire ($\sum_i p_i g(x_i) = \mathit{constante}$) peut se résoudre en minimisant une énergie libre qui n'est autre qu'une somme pondérée de l'entropie et du terme intervenant dans la contrainte.

A haute température, les probabilités p_i sont toutes égales, les états sont alors tous équiprobables indépendamment des énergies qui les caractérisent. A basse température, les états de faible énergie sont les plus probables. D'autre part, en substituant ces probabilités p_i dans l'expression de l'énergie libre, on a :

$$
\begin{aligned}
E_l^* &= \sum_{i=1}^{n} p_i E_i + T \sum_{i=1}^{n} p_i \log p_i \\
&= \sum_{i=1}^{n} p_i E_i + T \sum_{i=1}^{n} p_i \log e^{-E_i/T} - T \sum_{i=1}^{n} p_i \log \left(\sum_{j=1}^{j=n} e^{-E_j/T} \right) \\
&= \sum_{i=1}^{n} p_i E_i - \sum_{i=1}^{n} p_i E_i - T \log \left(\sum_{j=1}^{j=n} e^{-E_j/T} \right) \\
&= -T \log \left(\sum_{i=1}^{n} e^{-E_i/T} \right) = -T \log Z(T)
\end{aligned}
$$

d'où l'expression équivalente de l'énergie libre

$$E_l^* = -T \log Z(T) \tag{2.6.17}$$

Tout le calcul précédent suppose que les énergies E_i sont des constantes. Si tel n'est pas le cas, c'est-à-dire si les énergies E_i dépendent d'un paramètre y inconnu, on a alors dans ce cas :

$$E_l = E_l(p_1, ..., p_n, y) = \sum_{i=1}^{n} p_i E_i(y) - TS, \tag{2.6.18}$$

et on se retrouve face au problème suivant :

$$
\left\{
\begin{array}{l}
E_l^* = \min\limits_{p_1,\dots,p_n,y} \left(\sum\limits_{i=1}^{n} p_i E_i(y) + T \sum\limits_{i=1}^{n} p_i \log p_i \right) \\
\\
\text{avec} \quad \sum\limits_{i=1}^{n} p_i = 1.
\end{array}
\right. \tag{2.6.19}
$$

La résolution du problème 2.6.19 se fait en deux étapes, d'abord en fixant la variable y et en minimisant par rapport aux probabilités p_i, ce qui donne :

$$\begin{cases} E_l^*(y) = \min\limits_{p_1,...,p_n,y} E_l(p_1,...,p_n,y) = -T\log Z(T,y) \\ \\ \text{avec} \quad Z(T,y) = \sum\limits_{i=1}^{n} e^{-\frac{1}{T}E_i(y)}. \end{cases} \tag{2.6.20}$$

Le minimum de $E_l(p_1,...,p_n,y)$ est obtenu ensuite en minimisant $E_l^*(y)$ par rapport à la variable y.

Remarque 2.6.1
Si $P = (p_1,...,p_n)$ est la vraie loi que l'on cherche à estimer et si $P^y = (p_1^y,...,p_n^y)$ est la loi obtenue en minimisant l'énergie libre pour y fixé (selon 2.6.19), soit

$$\forall i, \quad 1 \le i \le n, \quad p_i^y = \frac{e^{-E_i(y)/T}}{\sum_{j=1}^n e^{-E_j(y)/T}},$$

alors, on peut écrire la distance de Kullback entre P et P^y, et on a

$$\begin{aligned} K(P,P^y) &= \sum_{i=1}^n p_i \log\frac{p_i}{p_i^y} \\ &= E_l/T + \log Z(T,y), \end{aligned}$$

avec

$$Z(T,y) = \sum_{i=1}^n e^{-\frac{E_i(y)}{T}}.$$

Comme $K(P,P_y) \ge 0$, on en déduit que

$$E_l \ge -T\log Z(T,y) = E_l^*(y),$$

ce qui est bien cohérent avec le résultat précédent puisque $E_l^*(y)$ est bien un minimum.

2.7 Application à la classification

2.7.1 Estimation

On va appliquer les résultats précédents au cas de la classification de données. Pour cela, on suppose qu'on dispose d'un ensemble de données, $\Xi = \{x_1,...,x_n\} \subset \mathbb{R}^d$, qu'on veut partitionner en m classes, avec $m \le n$. Nous supposons en premier lieu que les données sont des éléments de $\mathbb{R}$; la généralisation dans $\mathbb{R}^d$ est sans difficulté.

Soit y un point de $\mathbb{R}$, on cherche à calculer ce point de telle façon qu'on puisse le considérer comme un centre de classe de tout l'ensemble Ξ. Si la distribution de probabilité sur Ξ était connue, une bonne valeur de y serait l'espérance de Ξ qui réalise le minimum de la dispersion de Ξ. Dans notre cas cette distribution est inconnue, et il faut donc l'estimer. On suppose de plus qu'on ne dispose pas d'information suffisante permettant de choisir un modèle statistique sur Ξ. On se propose donc d'estimer simultanément cette distribution de probabilité P et le point y.

On pose $E_i(y) = (x_i - y)^2$, $E(y) = \sum_{i=1}^n p_i(x_i - y)^2$. On prend comme température $T > 0$ et on pose $\lambda = \frac{1}{T}$ comme plus haut. On estime alors $p_1, ..., p_n$ (dépendant de y) en minimisant l'énergie libre à y fixé

$$E_l(p_1, ..., p_n, y) = \sum_{i=1}^n p_i(x_i - y)^2 - TS.$$

Puis on minimise sur y. Les calculs sont menés dans la section suivante.

2.7.2 Recherche du centre y

D'après ce qui précède, on obtient à y fixé

$$\forall i, \quad 1 \leq i \leq n, \quad p_i^y = \frac{e^{-\lambda(x_i-y)^2}}{\sum_{j=1}^n e^{-\lambda(x_j-y)^2}},$$

et il reste à minimiser

$$E_l^*(y) = -\frac{\log Z(y,\lambda)}{\lambda} = -\frac{\log\left(\sum_{j=1}^n e^{-\lambda(x_j-y)^2}\right)}{\lambda}.$$

Cela revient à chercher les minima de la fonction F_λ, définie par :

$$F_\lambda(y) = \lambda E_l^*(y) = -\log Z(y,\lambda) = -\log\left(\sum_{i=1}^n e^{-\lambda(x_i-y)^2}\right).$$

On calcule donc les zéros de sa dérivée. Or

$$F'_\lambda(y) = -2\lambda \frac{\sum_{i=1}^n (x_i - y)e^{-\lambda(x_i-y)^2}}{\sum_{i=1}^n e^{-\lambda(x_i-y)^2}}$$

$$F'_\lambda(y) = 0 \quad \Longrightarrow \quad y = \frac{\sum_{i=1}^n x_i e^{-\lambda(x_i-y)^2}}{\sum_{i=1}^n e^{-\lambda(x_i-y)^2}} = \sum_{i=1}^n p_i^y x_i \tag{2.7.21}$$

Remarquons que y est le centre de gravité des x_i, pondérés par les p_i^y.

En posant

$$f_\lambda(y) = \frac{\sum_{i=1}^{n} x_i e^{-\lambda(x_i-y)^2}}{\sum_{i=1}^{n} e^{-\lambda(x_i-y)^2}},$$

L'estimation simultanée des p_i^y et de y revient donc à la solution de l'équation 2.7.21 et les centres de classes sont les points fixes de la fonction f_λ.

Il n'est pas possible de calculer explicitement les solutions y. Pour les calculer numériquement, on utilise un algorithme itératif du type $y_{t+1} = f_\lambda(y_t)$. Autrement dit, on écrit l'algorithme de descente du gradient pour la fonction F_λ :

$$y_{t+1} = y_t - \epsilon(t) F'_\lambda(y_t),$$

qui s'écrit, en prenant $\epsilon(t) = \frac{1}{2\lambda}$,

$$y_{t+1} = y_t + \frac{\sum_{i=1}^{n} (x_i - y_t) e^{-\lambda(x_i-y_t)^2}}{\sum_{i=1}^{n} e^{-\lambda(x_i-y_t)^2}} = f_\lambda(y_t). \qquad (2.7.22)$$

Le nombre et la nature des points fixes (attracteurs ou répulseurs) de cette fonction f_λ dépendent du paramètre λ qui, on a vu, joue le rôle de l'inverse d'une température.

Remarquons que pour $\lambda \gg 1$, on a $\forall x_i, f_\lambda(x_i) \approx x_i$ c'est-à-dire que tous les x_i sont approximativement des points fixes. Par contre pour $\lambda = 0, f_\lambda(y) = \frac{\sum_{i=1}^{n} x_i}{n}$ c'est-à-dire qu'on n'a qu'un seul point fixe qui est le barycentre de Ξ. Pour les valeurs intermédiaires de λ, il existe des points fixes, dépendant de λ, et dont le nombre est compris entre 1 et n. Nous étudions dans le chapitre suivant la fonction f_λ avant de donner le principe de l'algorithme de fusion proposé par Wong (Wong, 1993).

2.7.3 Étude de la fonction f_λ

Cas de deux points

Afin de mieux illustrer le comportement de la fonction f_λ, nous allons d'abord traiter le cas où $\Xi = \{x_1, x_2\} \subset \mathbb{R}$. Nous allons regarder le comportement et la nature des points fixes de f_λ suivant les valeurs de λ. La fonction f_λ est C^∞ et on suppose que $x_1 < x_2$. On a

$$\begin{aligned} f_\lambda(y) &= \frac{x_1 e^{-\lambda(x_1-y)^2} + x_2 e^{-\lambda(x_2-y)^2}}{e^{-\lambda(x_1-y)^2} + e^{-\lambda(x_2-y)^2}} \\ &= \frac{x_1}{1 + e^{-\lambda A(y)}} + \frac{x_2}{1 + e^{\lambda A(y)}} \end{aligned}$$

$$\text{avec} \quad A(y) = (x_2 - y)^2 - (x_1 - y)^2.$$

On a aussi :

$$\lim_{y\to-\infty} f_\lambda(y) = x_1 \quad , \quad \lim_{y\to+\infty} f_\lambda(y) = x_2$$

$$f_\lambda\left(\frac{x_1+x_2}{2}\right) = \frac{x_1+x_2}{2}$$

$$f'_\lambda(y) = \frac{2\lambda e^{\lambda A(y)}(x_2-x_1)^2}{\left(1+e^{\lambda A(y)}\right)^2} \quad , \quad f''_\lambda(y) = -\frac{2\lambda^2(x_2-x_1)^3 e^{\lambda A(y)}(1-e^{\lambda A(y)})}{\left(1+e^{-\lambda A(y)}\right)^2}.$$

On peut vérifier que le point (x_m, x_m) est centre de symétrie de la courbe représentative de f_λ.

Remarquons tout d'abord que $f'_\lambda(y)$ est toujours strictement positive, ce qui entraîne que f_λ est une fonction croissante et que $f_\lambda[\mathbb{R}] = [x_1, x_2]$. D'autre part on voit que $f'_\lambda(y)$ est croissante sur $]-\infty, x_m]$ de zéro à x_m et décroissante sur $[x_m, +\infty[$. Ce qui entraîne que $f'_\lambda(y)$ est strictement positive sur tout $\mathbb{R}$.

Ensuite, on voit que la suite y_t (définie par le premier terme y_0 et la relation de récurrence $y_{t+1} = f_\lambda(y_t)$ (eq.2.7.22)) est monotone. En effet d'après le théorème des accroissements finis, il existe θ réel, tel que :

$$y_{t+1} - y_t = f'_\lambda(\theta)(y_t - y_{t-1})$$

et $f'_\lambda(\theta) > 0$, pour tout θ. Comme cette suite est bornée entre x_1 et x_2, elle est nécessairement convergente vers l'un des points fixes. On connaît déjà un point fixe, x_m et qui vérifie :

$$f''_\lambda(x_m) = 0 \quad \text{et} \quad f'_\lambda(x_m) = \frac{\lambda(x_2-x_1)^2}{2}$$

Tous les points fixes sont solutions de

$$f_\lambda(y) = y,$$

ou encore si l'on pose $g_\lambda(y) = f_\lambda(y) - y$, de

$$g_\lambda(y) = 0.$$

On a $g'_\lambda(y) = f'_\lambda(y) - 1$, ce qui nous amène à distinguer deux cas suivant que

$$f'_\lambda(x_m) > 1 \quad \text{ou} \quad f'_\lambda(x_m) \leq 1.$$

On a immédiatement

$$f'_\lambda(x_m) > 1 \Leftrightarrow \lambda > \lambda_c,$$

$$\text{avec} \quad \lambda_c = \frac{2}{(x_2 - x_1)^2}.$$

Le nombre de points fixes dépend donc de λ et l'étude de la fonction f_λ est un problème de bifurcation par rapport à λ. D'où l'étude suivante qui distingue les cas $\lambda > \lambda_c$ et $\lambda \leq \lambda_c$

a) $\lambda > \lambda_c$ **:** Dans ce cas $f'_\lambda(x_m) > 1$. Donc il existe deux points (et seulement deux) $c_1 < x_m$ et $c_2 > x_m$ tels que $f'_\lambda(c_1) = f'_\lambda(c_2) = 1$. Si $g'_\lambda(y) = f'_\lambda(y) - 1$, on a immédiatement

$$g'_\lambda(y) < 0 \Leftrightarrow f'_\lambda(y) < 1 \Leftrightarrow -\infty < y < c_1$$

$\lim_{y \to -\infty} g_\lambda(y) = +\infty$, $g_\lambda(x_m) = 0$, d'où $g_\lambda(c_1) < 0$, d'où $\exists! x_1^* < c_1$, tel que $g_\lambda(x_1^*) = 0$, c'est-à-dire $f_\lambda(x_1^*) = x_1^*$.

On peut résumer tout cela dans le tableau de variation suivant (de $-\infty$ à x_m, le reste par symétrie)

	$-\infty$	x_1	x_1^*	c_1	x_m
$f'_\lambda(y)$	0	+	:	1	> 1
$g'_\lambda(y)$	-1	-	:	0	> 0
$g_\lambda(y)$	$+\infty$	$\searrow$	0	$< 0 \nearrow$	0
$f_\lambda(y)$	x_1	$\nearrow$	$x_1^* \nearrow$	$f_\lambda(c_1)$	x_m

De même, il existe un troisième point fixe x_2^*, symétrique de x_1^* par rapport à x_m.

On a donc trois points fixes $x_1^* < x_m < x_2^*$. Reste à étudier les limites qui ne peuvent être que l'un de ces points.

Tout d'abord si $y_0 \in]-\infty, x_1^*[$, $y_1 \in [x_1, x_1^*]$ (puisque f_λ est croissante) et pour tout t, $y_t \in [x_1, x_1^*]$. Alors

$$y_1 - y_0 = f_\lambda(y_0) - y_0 = g_\lambda(y_0) > 0,$$

la suite y_t est alors croissante, majorée par x_1^* et donc converge vers x_1^*.

Si $y_0 \in]x_1^*, x_m[$, alors $y_1 \in [x_1^*, x_m[$ (puisque f_λ est croissante) et pour tout t, $y_t \in [x_1^*, x_m[$. Alors $y_1 - y_0 < 0$, la suite y_t est décroissante, minorée par x_1^*, et donc converge vers x_1^*. De même si $y_0 > x_m$, la suite y_t converge vers x_2^*.

En conclusion si $\lambda > \lambda_c$, on a deux points fixes attracteurs x_1^* et x_2^* et un point répulsif x_m. On sait que, quand λ est grand, x_1^* est proche de x_1, (x_2^* est proche de x_2).

b) $\lambda \leq \lambda_c$ Dans ce cas $f'_\lambda(x_m) \leq 1$ et $\forall y \in \mathbb{R}$, $f'_\lambda(y) \leq 1$. Comme $g'_\lambda(y) \leq 0$ pour tout y, g_λ est décroissante et s'annule seulement en x_m. On peut résumer tout cela dans le tableau de variation suivant (de $-\infty$ à x_m, le reste par symétrie).

	$-\infty$		x_m
$f'_\lambda(y)$	0	+	< 1
$g'_\lambda(y)$	-1	-	< 0
$g_\lambda(y)$	$+\infty$	$\searrow$	0
$f_\lambda(y)$	x_1	$\nearrow$	x_m

Il y a donc dans ce cas un seul point fixe x_m.

On sait que la suite y_t est monotone.

Si $y_0 \in]-\infty, x_m[$, $y_1 = f_\lambda(y_0) = \in]x_1, x_m]$ et pour tout t, $y_t \in]x_1, x_m[$.

Comme $y_1 - y_0 > 0$, la suite y_t est croissante, majorée par x_m. Elle converge donc vers x_m. De même si $y_0 \in]x_m, +\infty[$, la suite est décroissante et converge vers x_m.

On a donc montré que si λ est suffisamment grand, ($\lambda > \lambda_c$) il y a 2 points fixes, solutions de l'équation 2.7.22 qui sont donc deux centres de classes possibles x_1^* et x_2^* proches des points eux-mêmes x_1 et x_2. Si λ est «petit» ; l'équation 2.7.22 n'a qu'une solution x_m qui est le centre de classe.

Cas de n points

Dans le cas où on a n points, $x_1 < x_2 < ... < x_i < ... < x_n$, on se ramène au cas précédent, en considérant les deux points x_i et x_j qui sont les plus proches. On peut alors, choisir $\epsilon > 0$ tel que sur l'intervalle $[x_1 - \epsilon, x_2 + \epsilon]$ on ait :

$$f_\lambda(y) \approx \frac{x_i e^{-\lambda(x_i - y)^2} + x_j e^{-\lambda(x_j - y)^2}}{e^{-\lambda(x_i - y)^2} + e^{-\lambda(x_j - y)^2}},$$

puisque les autres termes en exponentiel deviennent négligeables. La première confusion de points fixes concernera, en premier lieu, les points x_i^* et x_j^* proches des points x_i et x_j. Ces deux points fixes (x_i^* et x_j^*) vont se regrouper en un seul point fixe $\frac{x_i + x_j}{2}$ qui représentera $\{x_i, x_j\}$. On est alors ramené à un problème à $n - 1$ points et ainsi de suite en faisant décroître λ, on rapproche les points fixes de la fonction f_λ jusqu'à ce qu'il n'en reste qu'un seul point. Le cas où tous les points sont équidistants se traite en une seule fusion qui donne le barycentre de tous les points.

EXEMPLE 2.7.1 Voici un premier exemple simple avec deux points $x_1 = 2$ et $x_2 = 8$. Les quatre figures (2.1 et 2.4) illustrent le comportement de la fonction f_λ pour $\lambda = 4.5$ (fig 2.1), $\lambda = 0.08$ (fig 2.2), $\lambda = 0.06$ (valeur critique, fig 2.3) et $\lambda = 0$ (fig 2.4).

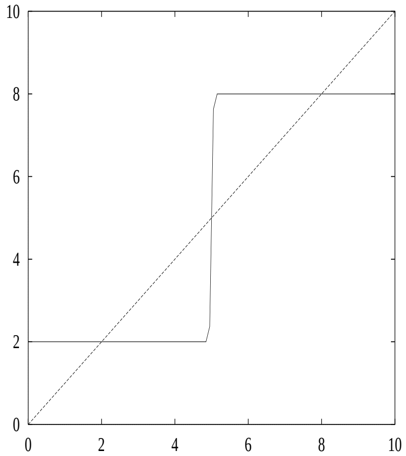


FIG. 2.1 - $\lambda = 4.5$

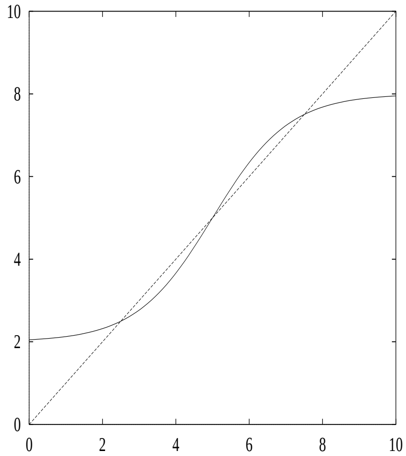


FIG. 2.2 - $\lambda = 0.08$

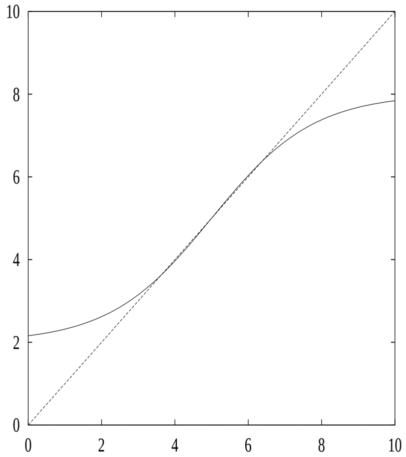


FIG. 2.3 - $\lambda = 0.06$ *valeur critique*

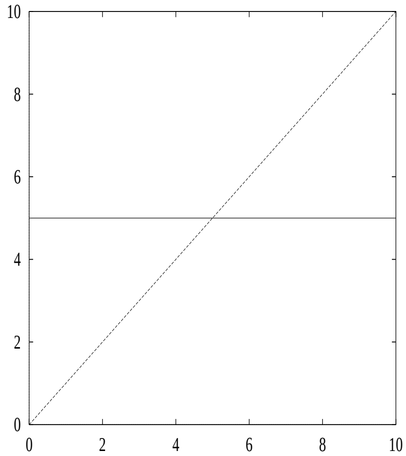


FIG. 2.4 - $\lambda = 0$

A basse température correspondant à une grande valeur de λ ($\lambda = 4.5$ (fig 2.1)), la courbe f_λ est approximativement en escalier, et elle possède trois points fixes $x_1^* \approx 2$, $x_2^* \approx 8$ et $x_m = 5$. Pour $\lambda = 0.08$, il y a toujours trois points fixes, avec $x_1^* > 2$ et $x_2^* < 8$; ces deux points fixes se rapprochent de $x_m = 5$ (fig 2.2). Dès que $\lambda = \lambda_c = 0.06$, c'est-à-dire la valeur critique, la courbe f_λ est de pente 1 au point $x_m = 5$ qui est le seul point fixe qui reste (fig 2.3). Quand $\lambda = 0$ (fig 2.4), la courbe est constante et vaut x_m. Les variations f_λ correspondent à un étirement de la courbe, où disparaissent progressivement les oscillations jusqu'à ce que l'on obtienne approximativement une droite.

EXEMPLE 2.7.2 Ici on se donne l'ensemble $\Xi = \{1, 2, 3, 8, 10, 12\}$; il est réunion de deux sous-ensembles de points : $\Xi_1 = \{1, 2, 3\}$ et $\Xi_2 = \{8, 10, 12\}$ tels que les trois premiers points sont distants d'une unité, les trois autres sont distants de deux unités (fig 2.5). Quand λ vaut 0.8 on observe qu'il y a eu une première confusion de points fixes, qui a regroupé les points de Ξ_1 en leur moyenne $\bar{x}_1 = 2$. Les points fixes correspondants aux points de Ξ_2 ne sont pas encore fusionnés (fig 2.6).

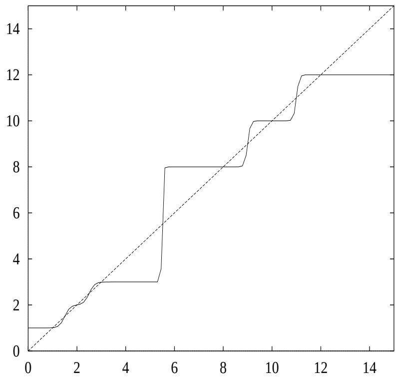


FIG. 2.5 - *courbe f_λ pour $\lambda = 4.5$*

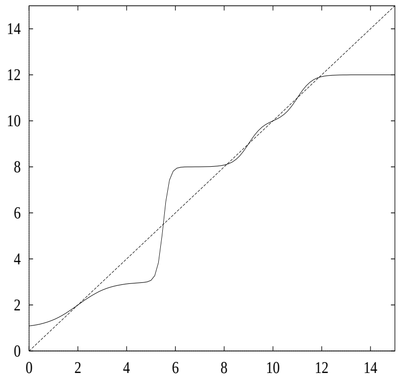


FIG. 2.6 - *courbe f_λ pour $\lambda = 0.8$*

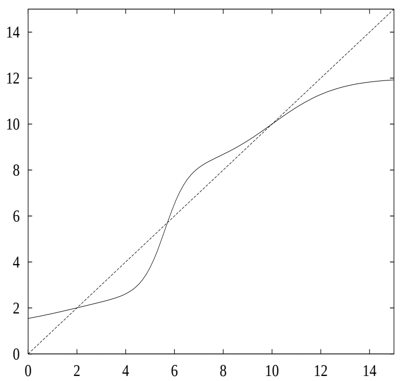


FIG. 2.7 - *courbe f_λ pour $\lambda = 0.2$*

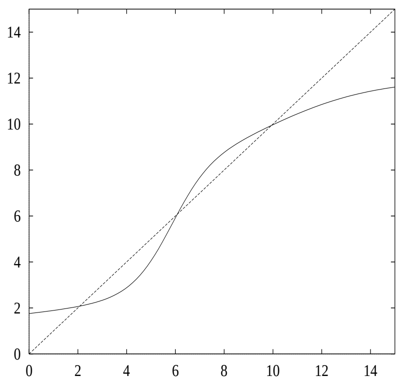


FIG. 2.8 - *courbe f_λ pour $\lambda = 0.1$*

Les propriétés de la fonction f_λ et de ses points fixes permettent de définir un algorithme de classification avec fusion progressive des centres de classes, qui a été proposé par Wong dans (Wong, 1993). Pour λ très grand, on part des n points x_i, eux-mêmes considérés comme les centres de n classes réduites à un point. Pour chacune des valeurs de λ (qui décroît), on détermine les centres de classes dont le nombre diminue en même temps que λ, jusqu'à ce que pour λ proche de 0, on obtienne une seule classe de centre $\frac{1}{n}\sum_i^n x_i$. On peut stopper le processus à tout instant pour obtenir un nombre de classes fixé à l'avance, ou bien chercher les «meilleurs» partitions, en un sens que l'on précisera dans le dernier paragraphe.

2.7.4 Définition de l'algorithme de fusion (Wong. Y. F.)

L'architecture de l'algorithme est composée d'un algorithme principal qui gère le paramètre de température λ et pour chaque valeur de ce paramètre λ il y a autant d'algorithmes qu'il y a de points fixes pour la fonction f_λ. L'algorithme principal ne fait que décroître le paramètre λ qu'on initialise par une grande valeur. Les sous-algorithmes sont tous identiques et permettent le calcul itératif des centres de classes. On adopte les notations suivantes :

t	compteur de l'algorithme principal
λ_t	paramètre de fusion à l'instant t
m_t	nombre de classes à l'instant t
j	indice de centre de classe $1 \leq j \leq m_t$
s	compteur pour les sous algorithmes avec $0 \leq s \leq S$
S	nombre maximum d'itération des sous-algorithmes

A l'instant t, il y a m_t centres de classes. Pour tous les centres de classes indicés par j tel que $1 \leq j \leq m_t$, on déplace leurs valeurs vers les points fixes de la fonction f_{λ_t}, soit :

étape 0 : INITIALISATION

$t = 1$
λ_1 grand
$m_1 = n$ nombre de classes initiales
$y^0_{1,S} = x_1, ..., y^0_{n,S} = x_n$

étape 1 : ALGORITHME PRINCIPAL (Tant que $m_t > m$)

1) sous algorithmes

$$\forall j,\ 1 \leq j \leq m_t \left\{ \begin{array}{l} s = 0 \quad y^t_{j,0} = y^{t-1}_{j,S} \\ \\ y^t_{j,s+1} = f_{\lambda_t}(y^t_{j,s}),\ 1 \leq s \leq S \end{array} \right.$$

2) fusion
si $y^t_{j,S} = y^t_{k,S}$, alors on supprime le centre $y^t_{k,S}$ et on renumérote les centres de 1 à $m_t - 1$

$m_t = m_t - 1$

3) décrémentation de λ_t

$$\lambda_{t+1} = \frac{\lambda_t}{\alpha(t)} \quad \alpha(t) > 1, \quad \text{(retour à l'étape 1)}$$

L'algorithme est mis en oeuvre tant qu'on n'a pas atteint le nombre de classes désiré. A la fin de l'algorithme on récupère les valeurs des centres de classes ainsi que les classes formées par la fusion des centres de classes depuis le début de l'algorithme. On obtient également la fonction qui définit le nombre de classes en fonction du temps et dont on regardera particulièrement les paliers afin de déterminer les nombres de classes pertinents. Mais avant de se pencher sur ce dernier point, on regarde ce que donne l'algorithme sur deux exemples réels.

2.7.5 Temps de fusion

On appelle temps d'une fusion, l'intervalle de temps que met l'algorithme entre deux fusions successives. C'est-à-dire si à l'instant t_1 il y a une fusion de classe et à l'instant t_2 une autre fusion de classe, tel que entre t_1 et t_2 il n'y a pas de fusion de classes, alors $t_2 - t_1$ est un temps de fusion. C'est le temps pour passer de m_{t_1} classes à m_{t_2} classes. Ainsi plus ce temps de fusion est grand plus la partition en m_{t_1} classe est significative.

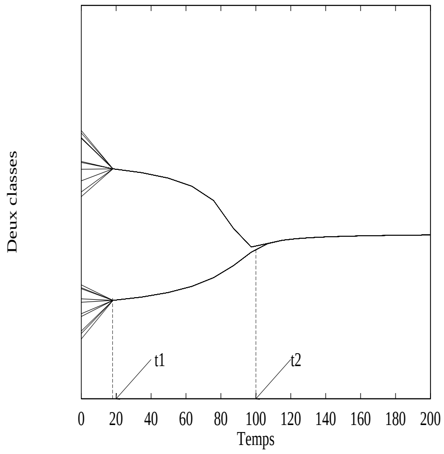


FIG. 2.9 - *Deux séries de points*

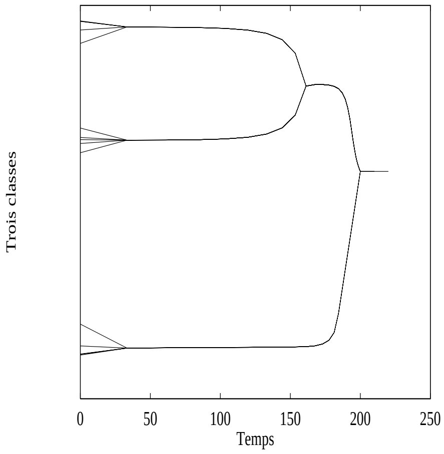


FIG. 2.10 - *Trois séries de points*

Dans l'exemple représenté par la figure 2.9, on a pris deux séries de 10 points tirés uniformément entre, 0 et 1 pour la première série, et 10 et 11 pour la deuxième série de points. La figure représente les trajectoires des centres de classes en fonction du temps. A l'instant 0 les centres de classes sont confondus avec les données initiales. Au cours des itérations les trajectoires se rapprochent et se confondent progressivement selon que leurs postions initiales étaient proches ou éloignées. Ainsi on peut constater que très rapidement les trajectoires des points de chaque série se confondent pour ne laisser place qu'à deux trajectoires qui correspondent aux centres de gravité de chaque série de points. Ces deux trajectoires vont mettre plus de temps pour se confondre, par rapport aux premières fusion. Cela coïncide bien avec le fait qu'on a pris deux classes bien séparables et suffisamment éloignées.

Dans le deuxième exemple (figure 2.10) ce sont trois séries de points, comprenant chacune cinq points, et tirés uniformément entre 0 et 1 pour la première, 10 et 11 pour la deuxième série et 15 et 16 pour la troisième. Comme pour l'exemple précédent les premières fusions de centres de classes s'opèrent très rapidement pour donner trois centres de classes qui vont rester suffisamment séparés pour fusionner ensuite en deux classes puis en une classe. On remarquera que le temps correspondant à la partition en trois classes est le plus important ce qui confirme la pertinence de cette partition entre trois classes.

2.8 Exemples

2.8.1 Les iris de Fisher

Les « *iris de Fisher* » sont les données de quatre variables que sont : la largeur des pétales, la longueur des pétales, la largeur des sépales et la longueur des sépales, mesurées sur 150 fleurs d'iris appartenant à trois espèces (*Setosa, Virginica, Versicolor*). On a mesuré ces quatre variables sur 50 fleurs pour chacune de ces espèces. On veut comparer les résultats obtenus par l'algorithme k-means, la classification hiérarchique et l'algorithme de fusion. Pour ce dernier on a pris $\lambda_0 = 0.3$ et $\alpha = 1.05$.

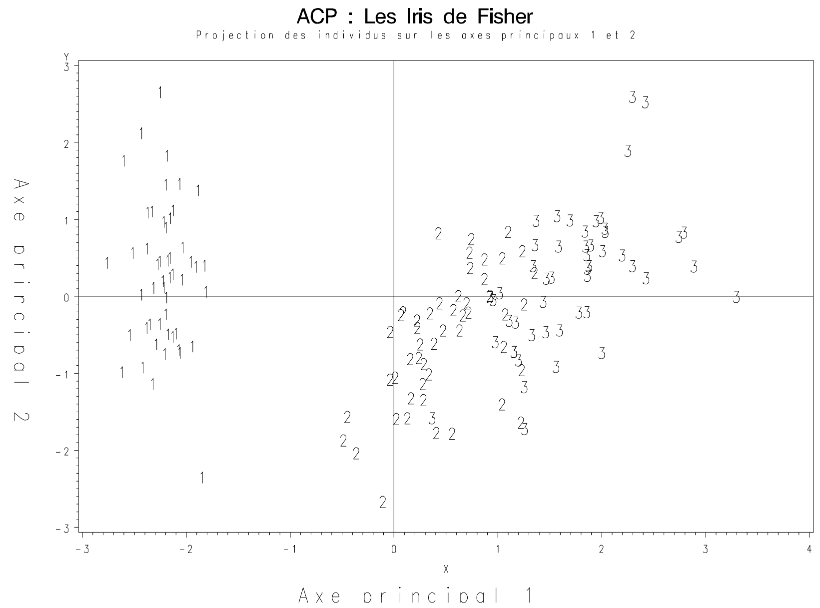


FIG. 2.11 - *Les Iris de Fisher*

Dans la figure 2.11 qui représente la projection 150 individus des trois espèces sur les deux premiers axes principaux montrent le recouvrement des espèces Virginica et Versivolor. Par contre, l'espèce Setosa est nettement linéairement séparable des deux autres espèces.

Le traitement par l'algorithme k-means et la classification hiérarchique ont été réalisées avec la procédure *fastclus* (pour k-means) et les procédures *cluster* et *tree* (pour la hiérarchique) du logiciel SAS. On regroupe les résultats avec ceux obtenus par l'algorithme de fusion.
Pour cette classification en trois groupes, on note que les trois algorithmes composent la classe 1 uniquement de l'espèce Setosa. Les différences entre les résultats de ces trois algorithmes concernent, ce qui était a priori prévisible, les deux autres espèces Versicolor et Virginica. Cependant on a plus de similitude entre l'algorithme de fusion et hiérarchique qu'avec k-means. En effet ces deux méthodes constituent une deuxième classe majoritairement avec l'espèce Virginica, alors que k-means constitue sa deuxième classe

	k-means			hiérarchique			fusion		
classes	Seto	Vers	Virg	Seto	Vers	Virg	Seto	Vers	Virg
1	50	0	0	50	0	0	50	0	0
2	0	2	36	0	15	49	0	23	49
3	0	48	14	0	35	1	0	27	1

TAB. 2.1 - *Les iris de Fisher en trois classes*

principalement avec l'espèce Versicolor. Il est connu que k-means dépend fortement de l'initialisation.

	k-means			hiérarchique			fusion		
classes	Seto	Vers	Virg	Seto	Vers	Virg	Seto	Vers	Virg
1	50	3	0	50	0	0	50	0	0
2	0	47	50	0	50	50	0	50	50

TAB. 2.2 - *Les iris de Fisher en deux classes.*

Dans le tableau 2.2 on voit que la première classe est constituée majoritairement (50 sur 53) de Setosa, alors que la deuxième classe est formée des espèces Versicolor et Virginica. On peut dire a posteriori, du fait que cette méthode dépend fortement de l'initialisation, que les trois Versicolor classées avec les Setosa devraient être classé dans la deuxième catégorie.

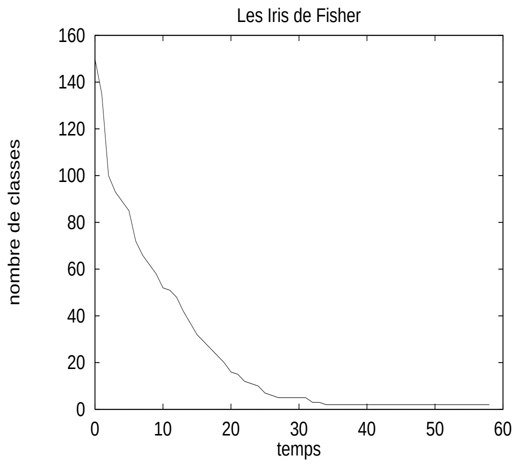


FIG. 2.12 - *Fusion des classes.*

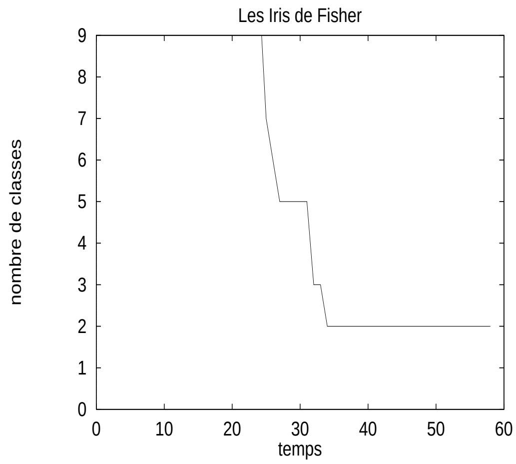


FIG. 2.13 - *Fusion des classes.*

La figure 2.12 représente la fusion des classes en fonction du temps, et on remarquera l'allure exponentielle de cette courbe. La figure 2.13 est un zoom sur la partie basse de la courbe 2.12 quand il ne reste que 9 classes. On remarque alors deux paliers qui se situent quand $m = 5$ puis quand $m = 2$. Ce dernier palier est plus long que le précédent et entre les deux on a un petit palier quand $m = 3$. Conclusion : une partition en deux classes est la plus pertinente.

2.8.2 Les arrondissements de Paris

Il s'agit dans cet exemple d'un extrait, concernant les vingt arrondissements de Paris, du fichier immobilier des 256 communes d'Ile de France. On a en effet 20 individus et 39 variables représentant les prix immobiliers et fonciers et des variables socio-économiques. Pour les variables prix on peut citer comme exemple le prix du sol (résidentiel et banal) le prix du logement neuf et ancien pour les catégories collectif et individuel. Les prix des bureaux anciens et neufs, les loyers de logement et de bureaux. Comme variable socio-économique on a le pourcentage de cadres, le taux de chômage, la densité par mètre au carré et le pourcentage d'étrangers.

On a traité par l'algorithme de fusion la classification des vingt arrondissements. On a pris $\lambda_0 = 1$ et $\alpha = 1.05$ ($\lambda_{t+1} = \frac{\lambda_t}{\alpha}$). Dans la figure 2.14 ci-dessous est représentée la fusion des classes en fonction du temps. Ainsi on peut constater les différents paliers à neuf classes, trois classes et deux classes. Nous comparerons ensuite la classification de l'algorithme de fusion à celle de la classification hiérarchique.

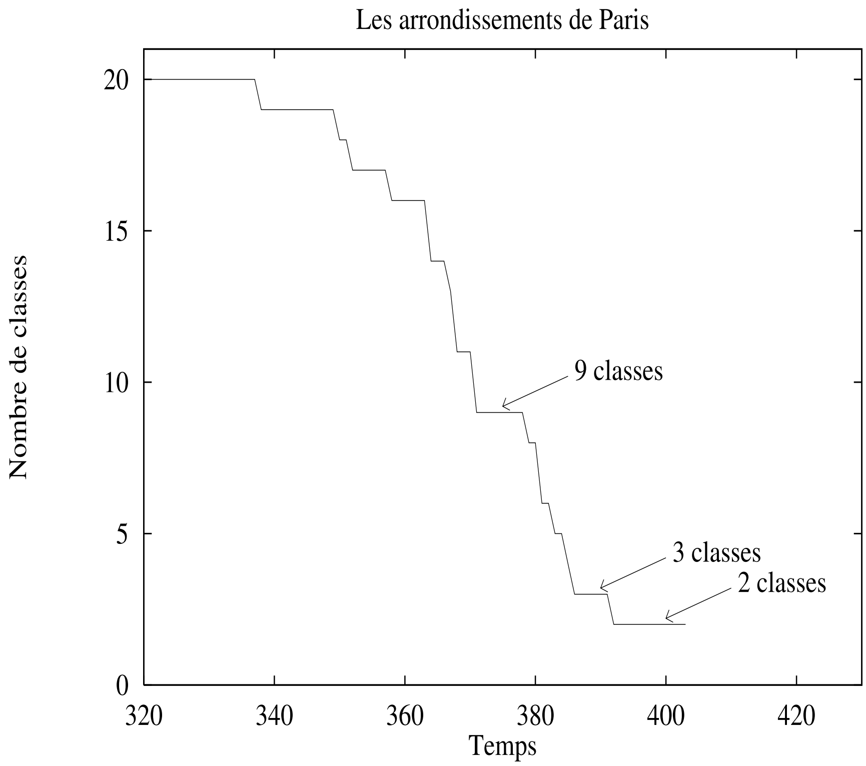


FIG. 2.14 - *Fusion des classes en fonction du temps*

On remarque que la valeur initiale λ_0 n'est pas facile à choisir car en effet on constate que son choix dans cet exemple est trop élevé puisque la première fusion n'arrive qu'après 330 itérations environ. Il n'y a pas de critère quantitatif pour choisir cette valeur initiale. La valeur $\lambda_0 = 0.3$ dans l'exemple des Iris de Fisher est plus adéquate.

individus	Les fusions successives																	
PARIS1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PARIS2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
PARIS3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	2	1	1	1	1
PARIS9	9	9	9	9	9	3	3	3	2	2	2	2	2	2	2	1	1	1	1
PARIS14	14	14	14	14	14	14	14	14	14	14	14	14	14	2	2	1	1	1	1
PARIS15	15	15	14	14	14	14	14	14	14	14	14	14	14	2	2	1	1	1	1
PARIS17	17	15	14	14	14	14	14	14	14	14	14	14	14	2	2	1	1	1	1
PARIS12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	2	1	1	1	1
PARIS10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	1
PARIS11	11	11	11	11	11	11	11	11	11	11	11	10	10	10	10	10	10	10	1
PARIS13	13	13	13	13	13	13	13	13	13	13	13	13	10	10	10	10	10	10	1
PARIS18	18	18	18	18	18	18	10	10	10	10	10	10	10	10	10	10	10	10	1
PARIS19	19	19	19	19	19	19	19	13	13	13	13	13	10	10	10	10	10	10	1
PARIS20	20	20	20	19	19	19	19	13	13	13	13	13	10	10	10	10	10	10	1
PARIS4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
PARIS5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4
PARIS6	6	6	6	6	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4
PARIS7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	5	4	4
PARIS8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7	7	5	4	4
PARIS16	16	16	16	16	16	16	16	16	16	16	7	7	7	7	7	7	5	4	4

TAB. 2.3 - *Classification par l'algorithme de fusion*

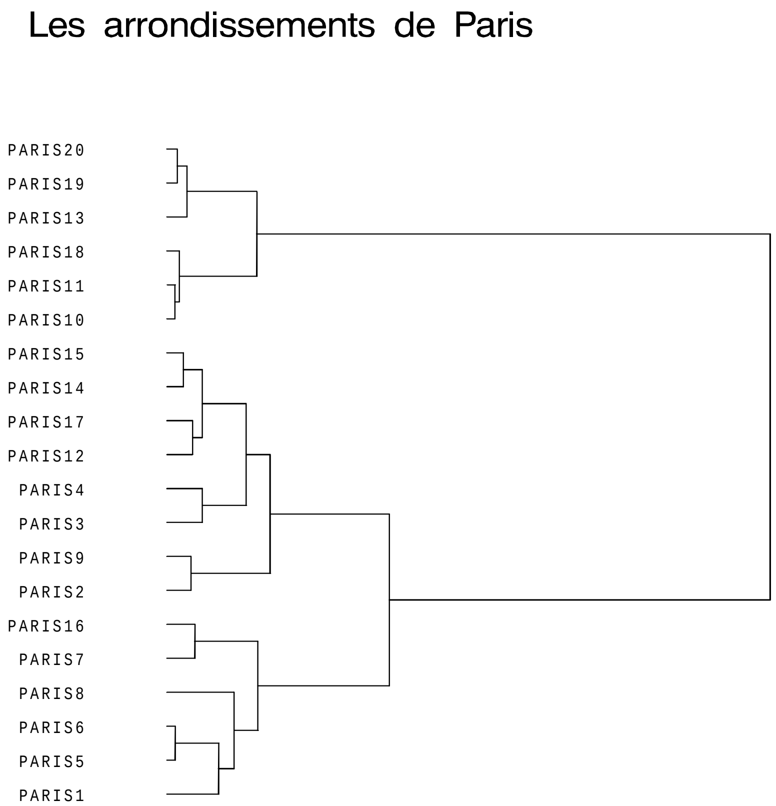


FIG. 2.15 - *Classification hiérarchique*

Du point de vue de la segmentation en trois classes des vingt arrondissements de Paris, les deux algorithmes donnent pratiquement la même partition. On retrouve pour l'algorithme de fusion une catégorie « banale » composée des arrondissements 10, 11, 18, 13, 19, et 20, une catégorie « moyenne » avec les arrondissements 1, 2, 3, 12, 14, 15 et 17 et une catégorie « supérieure » avec les arrondissements 16, 7, 8, 4, 5 et 6. Pour la classification hiérarchique (figure 2.15), on a juste une inversion de catégories pour le premier et le cinquième arrondissement. La classification en deux classes est assez distincte d'une méthode à l'autre, puisque la méthode par fusion regroupe les catégories « banale » et « moyenne » alors que la méthode hiérarchique regroupe les catégories « moyenne » et « supérieure ».

2.9 Conclusion

Il ressort de l'étude que nous avons faite de cet algorithme quelques remarques que nous décrivons ci-dessous :

L'algorithme de fusion est plus précis que l'algorithme k-means, en revanche il est beaucoup plus lent. Par rapport à la méthode hiérarchique, on peut dire que les erreurs sont comparables mais que l'algorithme de fusion doit être plus rapide sur les grandes bases de données.

Bien qu'il n'y ait pas de critère quantitatif pour le choix de λ_0, l'algorithme de fusion donne un moyen assez naturel pour choisir le nombre de classes à retenir. Ce qui est un bon atout pour cette méthode.

Chapitre 3

Méthodes classiques d'analyse des données

3.1 Introduction

L'*analyse des données* consiste à analyser les relations entre les différentes variables qui servent à décrire les individus, et à obtenir des représentations visuelles éclairant la structure interne de ces données. Dans un tableau de données, chaque observation ou individu est décrit par un vecteur ligne et chaque colonne correspond aux valeurs prises par une variable quantitative ou qualitative pour chacun des individus.

Deux techniques classiques sont particulièrement utilisées. « *L'analyse en composantes principales* » (**ACP**) traite des variables quantitatives et consiste à effectuer un changement de base orthonormée permettant de réduire le nombre utile des variables et de faire une représentation des individus en réduisant la dimension.

L'« *analyse factorielle des correspondances* » (**AFC**) peut être vue comme une analyse en composantes principales avec une pondération des individus et une métrique particulières. Elle est utilisée pour le traitement des tableaux de données comprenant deux variables qualitatives. « L'*analyse des correspondances multiples* » (**ACM**) est la généralisation de l'AFC au cas où l'on a plus de deux variables qualitatives.

Généralement ces tableaux de données qualitatives sont les résultats d'enquêtes ou de sondages, où les « individus » « répondent » à chaque question en choisissant une modalité, et une seule, parmi les modalités proposées pour cette question. On peut également faire une analyse des correspondances sur des variables continues (quantitatives) après discrétisation de leurs espaces des valeurs.

On se propose donc dans les paragraphes suivants d'exposer ces méthodes dites « linéaires ». On commence par faire un rappel suffisamment complet sur « *l'analyse en composantes principales* » qui sert de base pour la méthodologie de l'analyse des correspondances. On montre ensuite comment l'algorithme d'auto-organisation de Kohonen peut être utilisé pour obtenir une représentation simplifiée et informative des individus (algorithme **KACP**, chapitre 4) et également pour mettre en évidence les relations entre modalités de variables qualitatives (algorithme **KACM**, chapitre 6).

Nous rappelons dans cette section les principales notations et définitions usuelles. Nous nous sommes inspirés de la présentation faite par G.Saporta dans son livre *Probabilités, Analyse des données et Statistique*, (Saporta, 1990). On pourra se rapporter à cet ouvrage pour tout complément. On peut citer également le récent ouvrage de (Lebart et al., 1995) sur l'analyse multivariée.

3.2 Analyse en composantes principales (ACP)

3.2.1 Tableaux de données

Les observations de p variables sur n individus ou objets sont rassemblées dans une matrice X à n lignes et p colonnes, qui est appelée *matrice de données* ou encore *tableau de données.* Chaque ligne i de cette matrice représente les valeurs prises par l'individu i pour les p variables, et chaque colonne j désigne les n observations de la variable aléatoire réelle de numéro j.
Par abus de langage, on assimile une colonne de la matrice X à une variable. Chaque individu est vu comme un point de l'espace affine $\mathbb{R}^p$, mais que l'on confond avec un vecteur de $\mathbb{R}^p$. Chaque variable peut être vue comme un vecteur de $\mathbb{R}^n$. On note

$$X = (x_i^j)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \quad , \quad X = \begin{bmatrix} x_1^1 & & x_1^p \\ . & & . \\ x_i^1 & x_i^j & x_i^p \\ . & & . \\ x_n^1 & & x_n^p \end{bmatrix}$$

La variable j correspond au vecteur colonne de $\mathbb{R}^n$, $\quad x^j = \begin{pmatrix} x_1^j \\ . \\ . \\ . \\ x_n^j \end{pmatrix}$

L'individu i est le vecteur colonne x_i de $\mathbb{R}^p$ dont le transposé est $\quad x_i^t = \begin{pmatrix} x_i^1 & , . & . & ., & x_i^p \end{pmatrix}$
C'est le transposé de la i^{me} ligne de la matrice X.

Espace des individus - Matrice des poids

L'espace des individus est l'espace vectoriel $\mathbb{R}^p$. L'ensemble des individus constitue ce qu'on appelle un *nuage* de n points dans $\mathbb{R}^p$. Les individus peuvent être munis de poids qui caractérisent l'importance de chaque observation. C'est le cas, par exemple, des individus qui représentent des classes d'objets de tailles distinctes. On définit ainsi un vecteur de pondération sur l'ensemble des individus qu'on note $\pi = \begin{pmatrix} p_1 & \dots & p_n \end{pmatrix}$. Ce vecteur a toutes ses composantes positives et de somme 1, et peut être vu comme définissant une loi de probabilité.
La matrice des poids est alors la matrice diagonale D :

$$D = Diag\,\{p_1, p_2, ..., p_n\} = \begin{bmatrix} p_1 & & & \\ & p_2 & 0 & \\ & & & \\ & 0 & & \\ & & & p_n \end{bmatrix}$$

Si tous les individus ont la même importance, alors π est la loi uniforme $(1/n, ..., 1/n)$ et $D = \frac{1}{n} I_n$, où I_n est la matrice identité d'ordre n.

Individu moyen

On appelle *individu moyen* le centre de gravité g, vecteur de $\mathbb{R}^p$ dont les coordonnées sont les *moyennes* des variables éventuellement pondérées par π .

$$g^t = \begin{pmatrix} \bar{x}^1, & ... & , \bar{x}^j, & ... & \bar{x}^p \end{pmatrix} \quad \text{avec } \bar{x}^j = \sum_{i=1}^{i=n} p_i x_i^j$$

Si on note $\mathbf{1}_n$ le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1, alors

$$g = X^t D \mathbf{1}_n$$

La matrice X est dite *centrée* si $g = 0$, c'est-à-dire si et seulement si les p variables sont centrées.
La matrice Y définie par

$$Y = X - \mathbf{1}_n g^t = (I_n - \mathbf{1}_n \mathbf{1}_n^t D) X$$

est la *matrice des données centrées* (par colonne).

Matrice de variance-covariance - Dispersion

Alors que la position du nuage des individus est donnée par l'individu moyen g, la dispersion de ce nuage est mesurée par la matrice de variance-covariance empirique V des variables $(x^j)_{1 \leq j \leq p}$. On note

$$V = (v_{jk})_{\substack{1 \leq j \leq p \\ 1 \leq k \leq p}}$$

où

$$\begin{cases} v_{jj} = var(x^j) = \sum_{i=1}^{i=n} p_i (x_i^j)^2 - (\bar{x}^j)^2 = \sum_{i=1}^{i=n} p_i (x_i^j - \bar{x}^j)^2 \\ \\ v_{jk} = cov(x^j, x^k) = \sum_{i=1}^{i=n} p_i x_i^j x_i^k - \bar{x}^j \bar{x}^k = \sum_{i=1}^{i=n} p_i (x_i^j - \bar{x}^j)(x_i^k - \bar{x}^k) \quad \text{si} \qquad j \neq k \end{cases}$$

On note souvent σ_j^2 la variance empirique v_{jj}.
On remarque que

$$V = X^t DX - gg^t = Y^t DY$$

Si X est *centrée*, alors $V = X^t DX$. Si de plus, π est égale à la loi uniforme, alors $V = \sum_{i=1}^{i=n} \frac{1}{n} x_i x_i^t = \frac{1}{n} X^t X$. Cette égalité permet de calculer la matrice V d'une façon séquentielle.

On peut se ramener à des variables centrées et de variance 1, en considérant la matrice Z des données *centrées réduites* définie par :

$$Z = (z_i^j)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \qquad \text{où } z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma_j}$$

Espace des variables

En fait chaque variable est un vecteur de $\mathbb{R}^n$ et on peut identifier l'espace des variables avec l'espace vectoriel $\mathbb{R}^n$ qu'on munit de la métrique euclidienne définie par la matrice

$$D = Diag\{p_1, ..., p_n\}.$$

Alors les variances et covariances empiriques définies ci dessus peuvent être vues comme des normes et des produits scalaires des variables *si elles sont centrées*. On a :

$$\|x^j\|_D^2 = (x^j)^t D x^j = \sum_{i=1}^{i=n} p_i (x_i^j)^2 = var(x^j) = \sigma_j^2$$
$$< x^j, x^k >_D = (x^j)^t D x^k = \sum_{i=1}^{i=n} p_i x_i^j x_i^k = cov(x^j, x^k)$$

ce qui signifie que le produit scalaire pour D de deux variables (centrées) est égal à la covariance empirique des deux variables et que la norme pour D d'une variable est égale à son écart type empirique. Deux variables sont alors D-orthogonales si et seulement si elles sont non corrélées. Si la matrice des données est non seulement centrée mais réduite, alors le produit scalaire est exactement le cosinus de l'angle des deux variables.

Matrice de corrélation

La matrice des corrélations des p variables est notée R donnée par :

$$R = (r_{jk})_{\substack{1 \leq j \leq p \\ 1 \leq k \leq p}}$$

où

$$\begin{cases} r_{jj} = 1 \\ r_{jk} = corr(x^j, x^k) = \dfrac{cov(x^j, x^k)}{\sigma_j \sigma_k} \end{cases}$$

Elle mesure les relations des variables aléatoires observées pour les n individus.
En notant $M_\sigma = Diag\,\{\sigma_1, \ldots, \sigma_p\}$, alors $V = M_\sigma R M_\sigma$ et $R = M_{1/\sigma} V M_{1/\sigma}$ où $M_{1/\sigma} = M_\sigma^{-1} = Diag\,\{1/\sigma_1, \ldots, 1/\sigma_p\}$.

Si la matrice X est centrée réduite, $g = 0$ et $M_\sigma = I_p$, ce qui entraîne que $V = R$.

Si X est centrée réduite et si de plus $D = \frac{1}{n} I_n$, alors $V = R = \frac{1}{n} X^t X$. C'est le cas le plus utilisé en Analyse en Composantes Principales.

Métrique de l'espace des individus

On a vu que l'espace des individus est l'espace vectoriel $\mathbb{R}^p$. On munit cet espace d'une métrique définie par une matrice symétrique définie positive, de manière à pouvoir définir normes et distances.

Le choix de la métrique dans l'espace des individus est une tâche plutôt délicate car opter pour la métrique euclidienne usuelle -dont la matrice associée est la matrice identité dans $\mathbb{R}^p$- n'est pas toujours justifiée. En pratique, les variables sont généralement hétérogènes et possèdent des unités de mesures de grandeur différentes. On pense alors à une métrique M qui permette de pondérer les différentes composantes, de sorte que

$$\|x\|_M^2 = \sum_{j=1}^{j=p} a_j x_j^2,$$

ce qui correspond à une matrice diagonale

$$M = Diag\{a_1, .., a_p\}$$

Comme les variables sont pas nécessairement de même nature, et pour supprimer les effets tailles, on choisit en général de travailler sur des données centrées-réduites, ce qui équivaut à prendre la métrique définie par

$$M = M_{1/\sigma^2} = Diag\,\left\{1/\sigma_1^2, ..., 1/\sigma_p^2\right\}$$

où les termes de la diagonale sont les inverses des variances des p variables. Ainsi on réduit la variance de toutes les variables à l'unité, et chaque variable devient insensible au changement d'échelle. Cela est équivalent à travailler sur des données réduites.

Nous verrons que dans le cas de l'analyse factorielle des correspondances les individus sont des mesures de probabilités et la distance qu'on utilise est alors particulière. D'où l'intérêt de considérer le cas général où l'espace des individus est muni d'une métrique caractérisée par une matrice M symétrique définie positive.

3.2.2 Inertie d'un nuage de points

Comme nous l'avons rappelé, un ensemble de points est souvent appelé *nuage de points*. Ainsi le nuage des individus

$$\mathcal{N} = \{x_1, \dots, x_i, \dots, x_n\}$$

est un sous-ensemble de $\mathbb{R}^p$, qu'on suppose muni de la métrique définie par la matrice M. Les individus sont éventuellement pondérés par la matrice des poids D.

DÉFINITION 3.2.1
L'*inertie* d'un nuage de points par rapport à un point $\mathbf{a}$ est la somme des carrés des distances des points du nuage au point $\mathbf{a}$

$$I_{\mathbf{a}} = \sum_{x \in \mathcal{N}} p_i d_M^2(x, \mathbf{a})$$

DÉFINITION 3.2.2
L'*inertie totale* d'un nuage de points est l'inertie par rapport au centre de gravité du nuage.

$$I_g = \sum_{i=1}^{i=n} p_i d_M^2(x_i, g) = \sum_{i=1}^{i=n} p_i (x_i - g)^t M (x_i - g)$$

L'inertie totale traduit la dispersion du nuage de points autour du centre de gravité. Pour la suite, on note I_g l'inertie totale de $\mathcal{N}$.
Si $M = I_n$, l'inertie totale I_g n'est autre que la variance empirique totale du nuage de points.

PROPRIÉTÉ 3.2.1
L'inertie totale s'écrit $I_g = tr(MV) = tr(VM)$.

Démonstration : L'inertie est un scalaire, donc elle est égale à sa trace, et on a

$$\begin{array}{rcll} I_g & = & tr\left(\sum_{i=1}^{i=I} p_i (x_i - g)^t M (x_i - g)\right) & = tr\left(\sum_{i=1}^{i=I} p_i x_i M x_i - g^t M g\right) \\ & = & tr\left(\sum_{i=1}^{i=I} p_i M x_i x_i^t - M g g^t\right) & \text{car } tr(AB) = tr(BA) \\ & = & tr\left(M\left(\sum_{i=1}^{i=I} p_i x_i x_i^t - g g^t\right)\right) & = tr\left(M\left(X^t D X - g g^t\right)\right) \\ & = & tr(MV) & \end{array}$$

Enfin, comme M et V sont symétriques, on a bien le résultat.

Propriété 3.2.2 (Théorème d'Huygens)

$$I_a = I_g + \|g-a\|_M^2$$

Démonstration :

$$\begin{aligned}
I_a &= \sum_{i=1}^{i=n} p_i(x_i-a)^t M(x_i-a) \\
&= \sum_{i=1}^{i=n} p_i(x_i-g+g-a)^t M(x_i-g+g-a) \\
&= \sum_{i=1}^{i=n} p_i(x_i-g)^t M(x_i-g) + \|g-a\|_M^2 \\
&+(g-a)^t M \textstyle\sum_{i=1}^{i=n} p_i(x_i-g) + (\sum_{i=1}^{i=n} p_i(x_i-g))^t M(g-a) \\
&= \sum_{i=1}^{i=n} p_i(x_i-g)^t M(x_i-g) + \|g-a\|^2 \text{ car les deux autres termes sont nuls} \quad \square
\end{aligned}$$

De même, il existe une relation entre l'inertie totale et les distances entre tous les individus.

Propriété 3.2.3
La somme pondérée des écarts entre tous les individus est égale à deux fois l'inertie totale. Soit

$$2I_g = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} p_i p_j \|x_i - x_j\|_M^2$$

Démonstration :

$$\begin{aligned}
\sum_i \sum_j p_i p_j \|x_i-x_j\|_M^2 &= \sum_i \sum_j p_i p_j (x_i-x_j)^t M(x_i-x_j) \\
&= \sum_i \sum_j p_i p_j (x_i-g+g-x_j)^t M(x_i-g+g-x_j) \\
&= \sum_i \sum_j p_i p_j \big\{(x_i-g)^t M(x_i-g) + (g-x_j)^t M(x_i-g) \\
&\qquad +(x_i-g)^t M(g-x_j) + (g-x_j)^t M(g-x_j)\big\} \\
&= \sum_j p_j \sum_i p_i(x_i-g)^t M(x_i-g) + \sum_i \sum_j p_i p_j (g-x_j)^t M(x_i-g) \\
&+ \sum_i p_i \sum_j p_j (x_j-g)^t M(x_j-g) + \sum_i \sum_j p_i p_j (g-x_j)^t M(x_i-g) \\
&= 2I_g \quad \square
\end{aligned}$$

3.2.3 Projection de l'inertie, axes principaux

En ACP, on cherche à représenter les individus dans un sous-espace de plus petite dimension que l'espace de départ. On cherche alors des sous-espaces sur lesquels on va projeter le nuage de points tout en conservant au mieux la forme du nuage. Or après projection, les distances diminuent et de même pour l'inertie. Il faut donc rechercher des projecteurs tels que l'inertie projetée soit maximale.

A ce stade, on a besoin de définir l'*inertie projetée* sur un sous-espace vectoriel de $\mathbb{R}^p$:

DÉFINITION 3.2.3
Soient $x_1, ..., x_n$ des points de $\mathbb{R}^p$ et $y_1, ..., y_n$ leurs projections M-orthogonales sur un sous-espace vectoriel E de $\mathbb{R}^p$. L'inertie projetée sur E est l'inertie totale des y_i.

D'autre part, on peut montrer le théorème suivant

THÉORÈME 3.2.1 *Si E_k est un sous-espace de dim k maximisant l'inertie projetée, alors le sous-espace E_{k+1} de dim $k+1$ maximisant l'inertie projetée, peut s'écrire comme une somme directe $E_k \bigoplus F$ où $F \subset E^{\perp}$.*

On en déduit que ces espaces sont emboîtés et leur construction se fait de proche en proche à partir d'une droite vectorielle.

Dans le but de simplifier les calculs, nous supposerons dans la suite que la matrice des données X est centrée.

Soit a un vecteur directeur M-normé d'une droite Δ, et $P^{\Delta}(x_i) = x_i^{\Delta} a$ la projection M-orthogonale d'un individu x_i sur la droite Δ. Par définition $x_i - x_i^{\Delta} a$ est M-orthogonal à a et $\|a_M\| = a^t M a = 1$; soit :

$$\begin{array}{rcl} (x_i - x_i^{\Delta} a)^t M a & = & 0 \\ x_i^t M a - x_i^{\Delta} a^t M a & = & 0 \\ \text{et} \qquad x_i^{\Delta} & = & x_i^t M a \ (\text{car } a^t M a = 1) \end{array}$$

Donc le vecteur ligne, projection M-orthogonale de x_i sur la droite Δ s'écrit :

$$x_i^{\Delta} a = x_i^t M a a^t$$

et par conséquent, la matrice X^{Δ} de taille $n \times p$ des projections du nuage de points sur la droite Δ est :

$$X^{\Delta} = X M a a^t$$

La matrice X^{Δ} est centrée comme X, sa matrice de variance-covariance est donc donnée par $(X^{\Delta})^t D X^{\Delta}$ et son inertie totale est donnée d'après 3.2.1 par $tr((X^{\Delta})^t D X^{\Delta} M)$; soit :

$$\begin{array}{rcll} I^{\Delta} & = & tr(a(a^tMX^tDXMa)a^tM) & \\ & = & tr(a^tMX^tDXMa(a^tMa)) & \text{car } tr(AB) = tr(BA) \\ & = & a^tMX^tDXMa & \text{car } a^tMa = 1 \text{ et il s'agit d'un scalaire} \\ & = & a^tMVMa & \end{array}$$

La matrice MVM est appelée *matrice d'inertie.* Elle définit la forme quadratique qui à tout vecteur a associe l'inertie projetée sur la droite engendrée par ce vecteur. Pour trouver les directions maximisant l'inertie projetée, on cherche les maxima de I^{Δ} sous la contrainte $\|a\|_M^2 = 1$. Il faut donc résoudre le problème :

$$\begin{cases} \max\limits_{a \in \mathbb{R}^p} (a^tMVMa) \\ \\ a^tMa = 1 \end{cases}$$

On écrit alors le lagrangien $L(a)$ du problème qu'on dérive par rapport à a. Soit λ un multiplicateur de Lagrange réel,

$$L(a) = a^tMVMa - \lambda(a^tMa - 1)$$

$$L'(a) = 2MVMa - 2\lambda Ma$$

$$L'(a) = 0 \Longrightarrow MVMa = \lambda Ma$$

Comme M définit une métrique, elle est régulière (définie positive), et par conséquent

$$L'(a) = 0 \Longrightarrow VMa = \lambda a.$$

Le vecteur a est donc un vecteur propre de la matrice VM. L'inertie projetée I^{Δ} est égale à la valeur propre λ puisque :

$$I^{\Delta} = a^tMVMa = \lambda a^tMa = \lambda$$

Le vecteur directeur a qui correspond à une inertie projetée maximale est donc égal au vecteur propre associé à la plus grande valeur propre de VM.

Le problème se ramène donc à la diagonalisation de la matrice VM, $(p \times p)$. Cette matrice est M-symétrique, a toutes ses valeurs propres positives, et admet une base de vecteurs propres M-orthonormés.

Si la matrice VM est de rang $q \leq p$, on a alors q valeurs propres non nulles et qu'on range dans l'ordre décroissant. On les note :

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$$

On note $a_1, a_2, \ldots, a_q$ les q vecteurs propres M-orthogonaux associés.

On appelle transformation en composantes principales le changement de base de la base initiale supposée M-orthonormée vers la base M-orthonormée formée des q premiers vecteurs propres de VM, éventuellement complétés par une base du noyau de VM.

D'après le théorème 3.2.1, l'espace E_k de dimension k maximisant l'inertie projetée est alors engendré par les k vecteurs propres $a_1, a_2, \ldots, a_k$, associés aux k plus grandes valeurs propres.

Remarquons que si le rang de VM (égal au rang de V) est égal à $q < p$, l'inertie projetée sur l'espace E_q est égale à l'inertie totale.

Axes principaux

Les axes principaux sont les droites vectorielles engendrées par les vecteurs propres M-normés a de la matrice VM. Comme indiqué ci-dessus, si q est le rang de VM, on ne considère que les q premiers vecteurs propres, c'est-à-dire ceux qui correspondent aux q plus grandes valeurs propres de VM. Les axes sont numérotés selon l'ordre décroissant des valeurs propres, et notés comme les vecteurs propres $a_1, a_2, \ldots, a_q$.

Comme les axes principaux sont M-orthogonaux, l'inertie totale est la somme des inerties projetées sur chaque axe. Donc la somme de toute les valeurs propres est égale à l'inertie totale du nuage de points (l'inertie projetée sur l'axe est égale la valeur propre).

$$I_g = \lambda_1 + \lambda_2 + \ldots + \lambda_q$$

Le rapport

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_q}$$

est la proportion d'inertie expliquée par le k-ème axe principal. Il est très utile pour quantifier chaque axe principal.

Le rapport

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_q}$$

est l'inertie expliquée par les premiers axes. Il permet de définir le nombre d'axes à retenir.

Lorsque toutes les valeurs propres sont égales, le nuage de points est réparti d'une manière sphérique dans l'espace, et toutes les directions ont la même importance. Dans ce cas, l'ACP qui a pour objectif de réduire la dimension de l'espace des individus n'est pas une méthode appropriée.

Facteurs et composantes principales

Facteurs principaux : on appelle facteur principal f associé à l'axe principal engendré par le vecteur propre a, l'élément de $(R^p)^*$ espace dual de $\mathbb{R}^p$ (c'est-à-dire espace des formes linéaires sur $\mathbb{R}^p$) défini par le vecteur Ma appartenant à $\mathbb{R}^p$. C'est en fait le projecteur sur l'axe principal engendré par le vecteur a.

$$f : \left\{ \begin{array}{ccc} \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ x_i & \longrightarrow & x_i^t M a \end{array} \right.$$

On confond le facteur f et le vecteur Ma de $\mathbb{R}^p$. Les facteurs sont les vecteurs propres de la matrice MV associés aux mêmes valeurs propres que les vecteurs engendrant les axes principaux, puisque les valeurs propres de VM et de MV sont les mêmes.

$$MVf = MVMa = \lambda Ma = \lambda f$$

Rappelons que si un espace vectoriel E est muni d'une métrique M, l'espace dual E^* est alors muni de la métrique M^{-1}. Les facteurs principaux sont M^{-1}-orthonormés. En effet $f^t M^{-1} f = a^t M M^{-1} M a = a^t M a = 1$. De plus si f_1 et f_2 sont deux facteurs propres associés à deux axes principaux engendrés par deux vecteurs propres a_1 et a_2 M-orthonormés de MV, on a :

$$f_1^t M^{-1} f_2 = a_1^t M M^{-1} M a_2 = a_1^t M a_2 = 0.$$

On voit aussi que les facteurs sont aussi V-orthogonaux, puisque

$$f_1^t V f_2 = a_1^t M V M a_2 = \lambda_2 a_1^t M a_2 = 0 \text{ car } V M a_2 = \lambda_2 a_2$$

Composantes principales : On appelle composante principale c associée à l'axe principal défini par a, le vecteur colonne (de dimension n) composé des projections des individus sur cet axe. Les composantes principales sont les coordonnées des individus dans la nouvelle base. On a donc

$$c = Xf$$

C'est une variable générique qui est combinaison linéaire des variables x^j et dont la variance est égale à la valeur propre associée à a ou à f.

$$\begin{aligned} \|f\|_D^2 = Var(c) &= f^t X^t D X f \\ &= f^t V f \qquad (Vf = \lambda M^{-1} f) \\ &= \lambda f^t M_I^{-1} f \\ &= \lambda \end{aligned}$$

Reconstitution des données

Si le rang de V est $q \leq p$, la diagonalisation de la matrice MV fournit admet une base M^{-1}-orthonormée de vecteurs propres notés $(f_1, \ldots, f_p)$, dont les q premiers sont associés aux valeurs propres non nulles $\lambda_1, \ldots, \lambda_q$.

La matrice $\sum_{j=1}^{j=q} f_j f_j^t M^{-1}$ est égale à la matrice identité de $\mathbb{R}^p$. En effet, pour $1 \leq i \leq p$,

$$\left(\sum_{j=1}^{j=q} f_j f_j^t M^{-1}\right) f_i = \sum_{j=1}^{j=q} f_j f_j^t M^{-1} f_i = \sum_{j=1}^{j=q} \delta_{ij} f_j = f_i \text{ où } \delta_{ij} = \begin{cases} 0 \text{ si } i \neq j \\ 1 \text{ si } i = j \end{cases}$$

Si l'on note c_j la composante principale associée à f_j, on a

$$\begin{array}{l} X f_j = c_j \\ X f_j f_j^t M^{-1} = c_j f_j^t M^{-1} \\ X \left(\sum_{j=1}^{j=q} f_j f_j^t M^{-1}\right) = \sum_{j=1}^{j=q} c_j f_j^t M^{-1} \\ \\ \text{et donc } X = \sum_{j=1}^{j=q} c_j f_j^t M^{-1} \end{array}$$

On peut donc reconstituer la matrice des données (centrées) à partir des facteurs et composantes principaux. Ceci est équivalent aux formules de changement de bases orthonormée. De même, on a

$$\begin{aligned} MV f_j &= \lambda_j f_j \\ MV f_j f_j^t M^{-1} &= \lambda_j f_j f_j^t M^{-1} \\ MV &= \sum_{j=1}^{j=q} \lambda_j f_j f_j^t M^{-1} \end{aligned}$$

On peut aussi vérifier que $\sum_{j=1}^{j=q} a_j a_j^t M$ est égale à la matrice identité du fait que $a_i M a_j = \delta_{ij}$. Par suite

$$\begin{aligned} VM a_j &= \lambda a_j \\ VM a_j a_j^t M &= \lambda_j a_j a_j^t M \\ VM &= \sum_{j=1}^{j=q} \lambda_j a_j a_j^t M \end{aligned}$$

On obtient donc les résultats résumés suivants.

$$\boxed{\begin{array}{c} X = \sum_{j=1}^{j=q} c_j f_j^t M^{-1} \\ \\ MV = \sum_{j=1}^{j=q} \lambda_j f_j f_j^t M^{-1} \\ \\ VM = \sum_{j=1}^{j=q} \lambda_j a_j a_j^t M \end{array}}$$

3.2.4 Cas usuel : ACP sur données centrées réduites

Si on prend comme métrique sur $\mathbb{R}^p$ la matrice Identité, $MV = VM = V$, et on travaille directement sur la matrice de variance-covariance. Si les données sont de nature et d'échelle très différentes, il risque d'apparaître des effets de taille, et de plus l'analyse dépend fortement des unités de mesure. On préfère donc utiliser la métrique associée à la matrice des inverses des variances :

$$M = Diag\{\sigma_1, ..., \sigma_p\}$$

Il est équivalent de travailler avec des données centrées et réduites, et d'utiliser alors la métrique euclidienne usuelle sur $\mathbb{R}^p$. Avec les notations introduites précédemment, il vient

$$X = Z \quad M = I_p$$

En général on suppose aussi que tous les individus ont le même poids, et on choisit $D = \frac{1}{n} I_n$.
Ainsi

$$VM = MV = V = R = \frac{1}{n} X^t X.$$

Dans ce cas, les axes principaux et les facteurs principaux sont confondus. Ce sont les vecteurs propres de la matrice de corrélation R, rangés suivant l'ordre décroissant des valeurs propres. Un facteur principal f est donc un vecteur propre de $\frac{1}{n}X^tX$ et la composante principale c qui lui est associée est alors vecteur propre de la matrice $\frac{1}{n}XX^t$. En effet si $c = Xf$ et $\frac{1}{n}X^tXf = \lambda f$,

$$\frac{1}{n} XX^t c = \frac{1}{n} XX^t Xf = \lambda Xf = \lambda c$$

La composante principale c est une variable combinaison linéaire des variables initiales x^j, telle que la somme des corrélations carrées avec les variables initiales soit maximale.

L'analyse en composantes principales substitue donc aux variables originelles $x^1, x^2, \ldots, x^p$ (qui sont en général corrélées) des composantes principales combinaisons linéaires des variables initiales, qui sont non corrélées et de variances maximales.

On complète en général une analyse en composantes principales par le calcul des corrélations des variables initiales et des composantes principales. Pour une composante c associée à la valeur propre λ de la matrice de corrélation R, on a

$$Corr(x^j, c) = Corr(x^j, Xf) = \frac{\frac{1}{n}(x^j)^t Xf}{\sqrt{\lambda}}$$

puisque $Var(Xf) = Var(c) = \lambda$.

Le terme $\frac{1}{n}(x^j)^t X$ est la j-ième ligne de la matrice $\frac{1}{n}X^t X = R$, et comme $Rf = \lambda f$,

$$Corr(x^j, c) = \sqrt{\lambda} f_j$$

où f_j est la j-ième composante du vecteur f.

On peut donc terminer l'analyse en représentant les variables initiales dans les plans principaux successifs : axes 1,2, axes 1,3, axes 2,3, etc. Par exemple dans le plan (1,2), la variable x^j a pour coordonnées les corrélations $(Corr(x^j, c_1), Corr(x^j, c_2))$.

On représente également les individus dans ces différents plans en utilisant les nouvelles coordonnées après le changement de base.

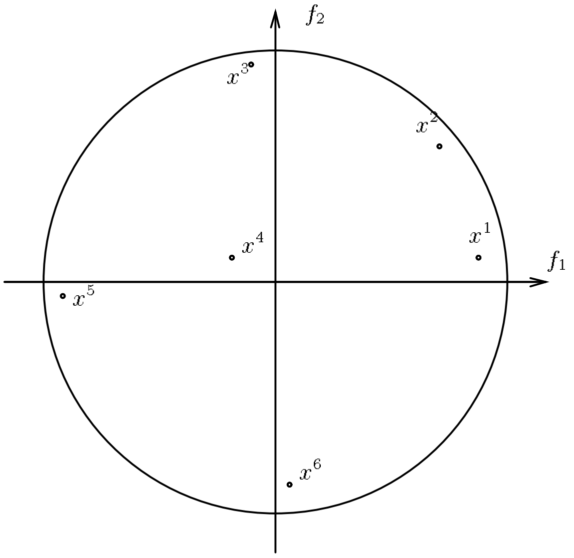


FIG. 3.1 - *Cercle de corrélations*

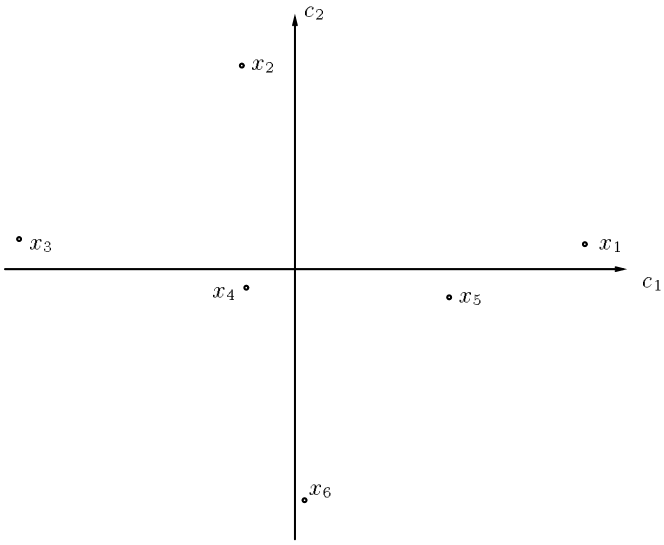


FIG. 3.2 - *Projection des individus*

La figure 3.1 représente les corrélations des variables par rapport aux facteurs f_1 et f_2. Plus les projections des variables sont proches du cercle de rayon 1, plus elles sont bien représentées dans ce plan de projection. C'est le cas, dans ce schéma, des variables x^1 et x^2 mais pas de la variable x^4. On dira également que le facteur f_1 oppose les variables x^1 et x^5, et que le facteur f_2 oppose les variables x^3 et x^6. La figure 3.2 représente la projection des individus sur le plan (1,2).

3.3 Analyse factorielle des correspondances (AFC)

Supposons qu'on ait un N-échantillon d'individus répondant à deux questions qu'on symbolise par X et Y. Les questions X et Y ont respectivement I et J modalités qualitatives.

3.3.1 Les tableaux de données

Il y a principalement deux façons de présenter les données d'un questionnaire à deux variables : le tableau disjonctif complet et le tableau de contingence. Le premier est un tableau contenant toute l'information concernant le questionnaire (individus et variables), le second par contre ne contient qu'un résumé sur les modalités des variables.

Le tableau disjonctif complet

C'est un tableau (matrice) K à N lignes et $I+J$ colonnes, dont les termes appartiennent à l'ensemble $\{0,1\}$. Le bloc des I premières colonnes correspond aux modalités de la variable X et les J colonnes restantes aux modalités de la variable Y. Chaque ligne représente un individu qui est caractérisé par une modalité et une seule pour chacune des variables. Nous affectons la valeur 1 à ces deux modalités et la valeur 0 aux autres modalités. D'où l'aspect disjonctif et complet du tableau.

K	X					Y				
	1	...	i	...	I	1	...	j	...	J
N Individus	0	0	0	1	0	0	0	1	0	0
	0	0	1			1				
	1	0	0	0	0	0	0	0	0	1
	1	0	0	0	0	0	1	0	0	0

TAB. 3.1 - *Tableau disjonctif complet*

C_{XY}	1	...	j	...	J	X
1			.			$n_{1.}$
.			.			.
.			.			.
i	.	.	n_{ij}	.	.	$n_{i.}$
.			.			.
.			.			.
I			.			$n_{I.}$
Y	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	N

TAB. 3.2 - *Tableau de contingence*

Tableau de Contingence

C'est une matrice à I lignes et J colonnes dont les termes représentent les nombres (effectifs) d'individus ayant choisi les différents couples de modalités des variables X et Y. Pour $1 \leq i \leq I$, $1 \leq j \leq J$ le terme n_{ij} représente le nombre d'individus ayant choisi la modalité i pour la variable X et la modalité j pour la variable Y. On note $C = (n_{ij})$ la table de contingence qui croise les variables X et Y.

3.3.2 Analyse d'un tableau de contingence

Soit un tableau de contingence C défini comme dans le paragraphe précédent.

$$C_{XY} = (n_{ij})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq J}}$$

On définit ensuite la matrice des fréquences

$$F_{XY} = \frac{1}{N} C_{XY}$$

On note f_{ij} le terme général de la matrice F_{XY}, il correspond à la probabilité empirique que la variable X prenne la valeur i et que la variable Y prenne la valeur j. On a :

$$P(X = i, Y = j) = f_{ij} = \frac{n_{ij}}{N}$$

$$F_{XY} = (f_{ij})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq J}}$$

La matrice F_{XY} correspond à la distribution de probabilité du couple (X, Y). A partir de ce tableau on calcule les probabilités marginales des variables X et Y qu'on désignera respectivement par les vecteurs F_X et F_Y. On note

$$P(X = i) = f_{i\cdot} = \sum_{j=1}^{j=J} f_{ij}, \quad 1 \leq i \leq I \quad , \quad P(Y = j) = f_{\cdot j} = \sum_{i=1}^{i=I} f_{ij}, \quad 1 \leq j \leq J$$

$$F_X = (f_{i\cdot})_{1 \leq i \leq I} \qquad F_Y = (f_{\cdot j})_{1 \leq j \leq J}$$

F_{XY}	1	...	j	...	J	F_X
1			.			$f_{1\cdot}$
.			.			.
.			.			.
i	.	.	f_{ij}	.	.	$f_{i\cdot}$
.			.			.
.			.			.
I			.			$f_{I\cdot}$
F_Y	$f_{\cdot 1}$	...	$f_{\cdot j}$	...	$f_{\cdot J}$	1

TAB. 3.3 - *Matrice des fréquences et marginales*

Les termes f_{ij} (resp n_{ij}) sont les fréquences observées (resp effectifs observés), les termes $f_{i\cdot}f_{\cdot j}$ (resp $n_{i\cdot}n_{\cdot j}/N$) sont les fréquences théoriques (resp effectifs théoriques) sous l'hypothèse d'indépendance des variables X et Y.

D'une manière générale, l'indépendance des variables X et Y s'écrit à partir de la distribution jointe et des distributions marginales. Dans notre cas la distribution jointe est décrite par la matrice F_{XY} et les distributions marginales par les vecteurs F_X et F_Y. Par conséquent, l'indépendance des variables X et Y est satisfaite si et seulement si

$$f_{ij} = f_{i\cdot} f_{\cdot j}\ ,\ 1 \leq i \leq I, \quad 1 \leq j \leq J \tag{3.3.1}$$

On a alors «*dépendance*» entre ces deux variables dès que certains termes f_{ij} de la matrice F_{XY} diffèrent du produit $f_{i\cdot} f_{\cdot j}$. D'une façon plus précise, on dit que les modalités i et j «*s'attirent*» (par rapport à l'hypothèse d'indépendance) si $f_{ij} > f_{i\cdot} f_{\cdot j}$ et «*se repoussent*» dans le cas contraire. L'égalité (3.3.1) peut être réécrite sous la forme suivante :

$$\forall i, \quad 1 \leq i \leq I\ , \quad \left(\frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}, \quad 1 \leq j \leq J \right) \tag{3.3.2}$$

qui signifie que toutes les lignes de la matrice F_{XY} sont proportionnelles à la probabilité marginale de Y. La quantité $\frac{f_{ij}}{f_{i\cdot}}$ est en fait $P(Y = j|X = i)$, c'est la proportion d'individus qui ont choisi la modalité j parmi tous les individus qui ont choisi la modalité i.

Ainsi l'égalité (3.3.2) signifie que

$$\forall j,\ 1 \leq j \leq J\ ,\ P(Y = j|X) = P(Y = j).$$

La variable X n'a aucune influence sur la variable Y. On adopte les deux notations suivantes :

$$f_j^i = \frac{f_{ij}}{f_{i\cdot}} \quad \text{et} \quad f_i^j = \frac{f_{ij}}{f_{\cdot j}}$$

et on introduit la notion de « *profils* » qui correspondent à des probabilités conditionnelles.

On appelle « *profil-ligne* » i la loi de Y sachant $(X = i)$, et on le désigne par f_Y^i. C'est le vecteur de $\mathbb{R}^J$ défini par :

$$f_Y^i = \{f_j^i, 1 \leq j \leq J\}$$

On désigne par F_Y^X la matrice de tous les profils lignes relatifs aux I modalités de la variable X. C'est une matrice à I lignes et J colonnes.

$$F_Y^X = \{f_Y^i, 1 \leq i \leq I\}$$

On représente dans le tableau 3.4 cette matrice F_Y^X. On a ajouté dans la dernière colonne de ce tableau les poids $f_{i\cdot}$ affectés à chaque profil-ligne. Le calcul du profil-ligne moyen g_I est explicité au paragraphe suivant.

F_Y^X	1	...	j	...	J	p_i
f_Y^1						$f_{1\cdot}$
.						
.						
f_Y^i	$\frac{f_{i1}}{f_{i\cdot}}$		$\frac{f_{ij}}{f_{i\cdot}}$		$\frac{f_{iJ}}{f_{i\cdot}}$	$f_{i\cdot}$
.						
.						
f_Y^I						$f_{I\cdot}$
g_I	$f_{\cdot 1}$		$f_{\cdot j}$		$f_{\cdot J}$	1

TAB. 3.4 - *Profils lignes*

F_X^Y	f_X^1	...	f_X^j	...	f_X^J	g_J
1			$\frac{f_{1j}}{f_{\cdot j}}$			$f_{1\cdot}$
.			.			.
.			.			.
i			$\frac{f_{ij}}{f_{\cdot j}}$			$f_{i\cdot}$
.			.			.
.			.			.
I			$\frac{f_{Ij}}{f_{\cdot j}}$			$f_{I\cdot}$
p_j	$f_{\cdot 1}$		$f_{\cdot j}$		$f_{\cdot J}$	1

TAB. 3.5 - *Profils colonnes*

Comme pour les I modalités de la variable X, on construit la matrice des profils colonnes correspondants aux J modalités de la variable Y.

Le « *profil-colonne* » j noté f_X^j est la loi de X sachant $(Y = j)$, c'est le vecteur de $\mathbb{R}^I$ défini par :

$$f_X^j = \{f_i^j, 1 \leq i \leq I\}.$$

La matrice des profils colonnes est notée F_X^Y

$$F_X^Y = \{f_X^j, 1 \leq j \leq J\}$$

On obtient ainsi deux nuages de points qu'on traitera d'une façon symétrique : un nuage de profils lignes dans $\mathbb{R}^J$ et un nuage de profils colonnes dans $\mathbb{R}^I$. L'AFC consiste à rechercher une typologie des profils lignes par rapport aux modalités colonnes et une typologie des profils colonnes par rapport aux modalités lignes. Pour cela, on fait simultanément deux ACP sur ces deux nuages avec des métriques et des pondérations symétriques.

Métrique du Chi 2

On appelle statistique du χ^2 d'un tableau de contingence la somme des carrés des écarts, rapportés aux effectifs théoriques, entre les effectifs observés et les effectifs théoriques.

$$\chi^2 = \sum_{ij} \frac{(Nf_{ij} - Nf_{i\cdot}f_{\cdot j})^2}{Nf_{i\cdot}f_{\cdot j}}$$

Nous verrons que ce terme est égal-à la constante N multiplicative près, à l'inertie du nuage des profils lignes (ou des profils colonnes, vu la symétrie du problème) par rapport à la métrique définie ci-dessous et qu'on appelle alors la «*métrique du* χ^2».

La ressemblance entre deux profils (lignes, colonnes) se mesure par la distance du χ^2 qui opère d'une manière symétrique sur les profils lignes et sur les profils colonnes :

DÉFINITION 3.3.1
On définit la distance du χ^2 entre deux profils lignes i et l par

$$d_{\chi^2}(f_Y^i, f_Y^l) = \sum_{j=1}^{j=J} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2$$

ainsi que la distance du χ^2 entre deux profils colonnes j et k par

$$d_{\chi^2}(f_X^j, f_X^k) = \sum_{i=1}^{i=I} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ik}}{f_{\cdot k}} \right)^2$$

C'est la distance euclidienne usuelle dont chaque terme quadratique est pondéré par l'inverse de la marginale de la variable correspondante. D'une façon analogue à l'ACP lorsqu'on utilise la matrice centrée réduite pour enlever l'effet taille, cette distance possède l'avantage de ne pas négliger les modalités rares.
La principale justification de l'utilisation de cette distance du Chi-deux est due particulièrement à la propriété dite «*d'équivalence distributionnelle*» : s'il existe deux colonnes proportionnelles dans le tableau C et si on remplace ces deux colonnes par la colonne égale à leur somme, alors les distances entre les profils lignes restent inchangées. Ceci découle de l'égalité ci-dessous.
Soient deux colonnes de C indicées par j et j'

$$(n_{ij})_{1 \leq i \leq I} \qquad (n_{ij'})_{1 \leq i \leq I}$$

telles que

$$\exists \alpha \in \mathbb{R} \;/\; n_{ij} = \alpha n_{ij'} \qquad 1 \leq i \leq I$$

On somme les deux colonnes en une seule

$$(n_{ij} + n_{ij'})_{1 \leq i \leq I}.$$

Si on calcule la distance du χ^2 entre deux profils lignes i et i', on s'aperçoit qu'on a deux termes correspondants aux colonnes j et j' de C qui sont remplacés par le terme ci-dessous relatif à cette nouvelle colonne. Ce dernier est en fait la somme des deux précédents. En effet :

$$\begin{aligned}
\frac{N}{n_{\cdot j} + n_{\cdot j'}} \left(\frac{n_{ij} + n_{ij'}}{n_{i\cdot}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'\cdot}} \right)^2 &= \frac{N}{n_{\cdot j'}} \left(\frac{n_{ij'}}{n_{i\cdot}} - \frac{n_{i'j'}}{n_{i'\cdot}} \right)^2 (1+\alpha) \\
&= \frac{N}{n_{\cdot j'}} \left(\frac{n_{ij'}}{n_{i\cdot}} - \frac{n_{i'j'}}{n_{i'\cdot}} \right)^2 + \alpha \frac{N}{n_{\cdot j'}} \left(\frac{n_{ij'}}{n_{i\cdot}} - \frac{n_{i'j'}}{n_{i'\cdot}} \right)^2 \\
&= \frac{N}{n_{\cdot j'}} \left(\frac{n_{ij'}}{n_{i\cdot}} - \frac{n_{i'j'}}{n_{i'\cdot}} \right)^2 + \frac{N}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2
\end{aligned}$$

Donc ce terme est égal à la somme des deux termes correspondants aux deux colonnes supprimées. On a évidemment le résultat équivalent dans le cas où l'on a des lignes proportionnelles.

Ces deux distances du χ^2 sur $\mathbb{R}^J$ et sur $\mathbb{R}^I$ sont des distances qui confèrent à leurs espaces respectifs une structure d'espace euclidien. On note par M_I et M_J les matrices des métriques correspondantes :

$$M_I = \begin{pmatrix} \frac{1}{f_{\cdot 1}} & & \\ & \frac{1}{f_{\cdot j}} & \\ & & \frac{1}{f_{\cdot J}} \end{pmatrix} \quad M_J = \begin{pmatrix} \frac{1}{f_{1\cdot}} & & \\ & \frac{1}{f_{i\cdot}} & \\ & & \frac{1}{f_{I\cdot}} \end{pmatrix}$$

Les nuages des profils lignes $\mathcal{N}_\mathcal{I}$ et profils colonnes $\mathcal{N}_\mathcal{J}$

Il est important de noter que les profils lignes sont considérées comme des points de l'espace affine $\mathbb{R}^J$. L'ensemble de ces I points est noté $\mathcal{N}_\mathcal{I}$, et on note $\mathcal{N}_\mathcal{J}$ l'ensemble des profils colonnes considérés comme des points de l'espace affine $\mathbb{R}^I$.

a) Support des profils : le nuage des profils lignes est une suite de I points de $\mathbb{R}^J$ dont le support est inclus dans un hyperplan affine H_I, en effet :

$$\forall i, \quad 1 \leq i \leq I \quad \sum_{j=1}^{j=J} f_j^i = \sum_{j=1}^{j=J} \frac{f_{ij}}{f_{i\cdot}} = 1$$

H_I est l'hyperplan affine d'équation $\sum_{i=1}^{i=J} x_i = 1$. De même le nuage $\mathcal{N}_J$ est contenu dans un hyperplan affine H_J d'équation $\sum_{j=1}^{j=I} x_j = 1$

b) Matrices des poids : Chaque profil-ligne i est affecté du poids $f_{i\cdot}$. La matrice des poids D_I est donc égale à la matrice M_J^{-1}. Chaque profil-colonne j est affecté du poids $f_{\cdot j}$ et la matrice des poids D_J est égale à M_I^{-1}.

c) Profils moyens : Le profil-ligne moyen g_I est par définition le barycentre des profils lignes affectés de leurs poids respectifs.

$$g_I = \sum_{i=1}^{i=I} f_{i\cdot} f_Y^i$$

dont les coordonnées sont :

$$g_I(j) = \sum_{i=1}^{i=I} f_{i\cdot} f_j^i = \sum_{i=1}^{i=I} f_{ij} = f_{\cdot j}, \quad 1 \leq j \leq J.$$

On a par conséquent :

$$g_I = F_Y$$

Rappelons $g_I = (F_Y^X)^t D_I \mathbf{1}_J^t = F_Y$ où $\mathbf{1}_J$ est le vecteur de $\mathbb{R}^J$ dont toutes les coordonnées sont égales à 1. Notons que g_I est bien un point de l'hyperplan H_I.

Factorisation des inerties des deux nuages

Dans ce paragraphe on va surtout appliquer les résultats obtenus sur l'ACP et qu'on commentera ensuite. Tout d'abord on regarde la relation entre l'inertie et la statistique du Chi-deux d'un tableau de contingence.

$$\begin{aligned} Inertie(\mathcal{N}_{\mathcal{I}}) &= \sum_{i=1}^{i=I} f_{i\cdot}(f_J^i - g_I)^t M_I (f_J^i - g_I) = \sum_{i=1}^{i=I} f_{i\cdot} \sum_{j=1}^{j=J} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 \\ &= \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \frac{\chi^2}{N}. \end{aligned}$$

De même on a le résultat analogue pour l'inertie du nuage des profils colonnes.

$$Inertie(\mathcal{N}_{\mathcal{J}}) = \frac{\chi^2}{N}$$

Le profil-ligne moyen g_I est un vecteur M_I-orthogonal à l'hyperplan affine H_I.

En effet soit

$$a = (a_j)_{1 \leq j \leq J}$$

un point appartenant à l'hyperplan affine H_I. On a g_I qui est M_I-orthogonal à $g_I - a$:

$$g_I^t M_I (a - g_I) = \sum_{j=1}^{j=J} \frac{f_{\cdot j}}{f_{\cdot j}} (a_j - f_{\cdot j}) = \sum_{j=1}^{j=J} a_j - 1 = 0$$

Cela traduit en fait que l'inertie projetée sur l'axe engendré par le vecteur g_I est nulle.

En se référant au chapitre sur l'ACP, on considère la matrice des données X qui est égale à la matrice des profils lignes

$$X = F_Y^X = M_J F_{XY},$$

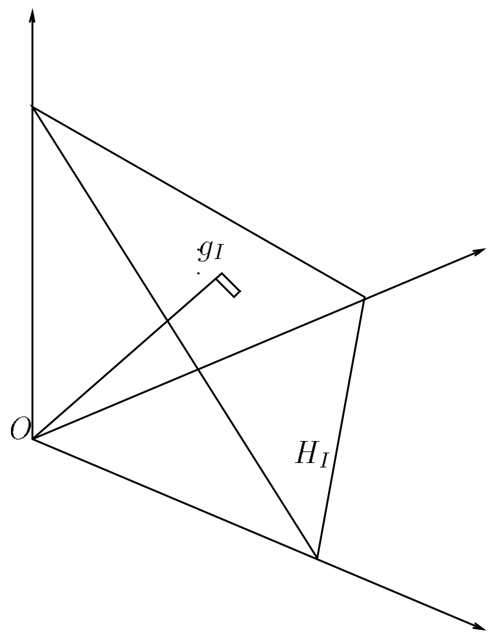


FIG. 3.3 - *Hyperplan des profils lignes*

la matrice des poids D qui est l'inverse de la matrice définissant la métrique dans l'espace des profils colonnes

$$D = M_J^{-1},$$

et finalement la métrique est donnée par M_I

$$M = M_I.$$

La matrice d'inertie est égale à $M_I V M_I$ où $V = X^t D X - g_I g_I^t$. Or on a déjà établi que la forme quadratique $a^t M_I V M_I a$ définit l'inertie projetée sur l'axe engendré par a. Or comme g_I est orthogonal à l'hyperplan H_I on a donc

$$g_I^t M_I V M_I g_I = 0$$

donc $M_I g_I = 0$, ce qui entraîne $V M_I g_I = 0$, c'est-à-dire que g_I est vecteur propre de la matrice $V M_I$ associé à la valeur propre zéro. C'est un axe principal trivial et il correspond au facteur principal

$$M_I g_I = \mathbf{1}_J,$$

le vecteur de $\mathbb{R}^J$ dont toutes les coordonnées sont égales à un.

D'autre part, soit a un axe principal c'est-à-dire un vecteur propre de la matrice $V M_I$, a est aussi vecteur propre de $X^t D X$:

$$\begin{aligned} X^t DX &= V + g_I g_I^t \\ X^t DXM_I a &= VM_I a + g_I g_I^t M_I a \\ X^t DXM_I a &= VM_I a = \lambda a \quad \text{car } g_I M_I a = 0 \end{aligned}$$

De plus on a

$$\begin{aligned} X^t DXM_I g_I &= VM_I g_I + g_I g_I^t M_I g_I \\ X^t DXM_I g_I &= g_I \quad \text{car } \|g_I\|_{M_I} = 1. \end{aligned}$$

g_I est donc vecteur propre de la matrice $X^t DXM_I$ associé à la valeur propre un. Ainsi les vecteurs propres de la matrice VM_I sont identiques à ceux de $X^t DX_I$. De ce fait on peut donc rechercher les vecteurs propres de $M_I X^t DX$ sans centrer la matrice des profils lignes.

En conclusion les axes principaux sont vecteurs propres de la matrice $X^t DXM_I$ distincts de g_I (on ne considère que les valeurs propres distinctes de zéro). Les facteurs principaux sont les vecteurs propres de $M_I X^t DX$ distincts du vecteur $\mathbf{1}_J$ (les valeurs propres sont distinctes de un). Les composantes principales sont les vecteurs propres de $XM_I X^t D$ et on a:

$$X^t DXM_I = F_{XY}^t M_J M_J^{-1} M_J F_{XY} M_I = F_{XY}^t M_J F_{XY} M_I$$

$$M_I X^t DX = M_I F_{XY}^t M_J M_J^{-1} M_J F_{XY} = M_I F_{XY}^t M_J F_{XY}$$

$$X^t M_I XD = M_J F_{XY} M_I F_{XY}^t M_J M_J^{-1} = M_J F_{XY}^t M_I F_{XY}^t$$

Pour les profils colonnes on a les résultats analogues en prenant comme matrice des données

$$X = (F_X^Y)^t = (F_{XY} M_I)^t = M_I F_{XY}^t$$

et comme matrice de poids et métrique respectivement

$$D = M_I^{-1} \quad M = M_J.$$

Ce qui correspond aux résultats finaux du tableau suivant.

Remarques

Dans les deux paragraphes suivants nous allons traiter le cas où le nombre des variables est supérieur à deux. Mais avant de regarder ce cas qu'on appelle dans la littérature « *analyse*

profils lignes	vecteurs propres de
axes	$F_{XY}^t M_J F_{XY} M_I$
Facteurs	$M_I F_{XY}^t M_J F_{XY}$
composantes	$M_J F_{XY} M_I F_{XY}^t$
profils colonnes	vecteurs propres de
axes	$F_{XY} M_I F_{XY}^t M_J$
Facteurs	$M_J F_{XY} M_I F_{XY}^t$
composantes	$M_I F_{XY}^t M_J F_{XY}$

des correspondances multiples », nous allons faire quelques remarques qui peuvent être utiles dans la suite.

Les matrices des profils que nous avons définies à partir des fréquences peuvent être calculées à l'aide des effectifs. On peut remarquer que

$$f_j^i = \frac{f_{ij}}{f_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}}$$

Les facteurs sont les vecteurs propres de la matrice MV, avec $V = X^t DX$ et où on a $X = F_Y^X$. Si on pose

$$Z_Y^X = (z_{ij}) \qquad \text{avec} \qquad z_{ij} = \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}},$$

alors $M = Z_Y^X D(Z_Y^X)^t$, ce qui revient à corriger la matrice des entrées et utiliser une ACP sur la nouvelle matrice avec une métrique euclidienne et une matrice des poids égale à D.

D'autre part, les facteurs principaux sont les vecteurs propres de la matrice $M_I F_{XY}^t M_J F_{XY}$ qui est égale, après avoir effectué les calculs, au produit $T^t T$ où

$$T = (t_{ij}) \qquad \text{et} \qquad t_{ij} = \frac{f_{ij}}{\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j}}}$$

On peut donc faire une ACP avec une pondération uniforme et une métrique euclidienne sur la matrice des données corrigées T. Nous utiliserons ces deux corrections pour définir l'algorithme neuronal KACM pour le traitement des variables qualitatives.

3.4 Analyse des correspondances multiples

Comme pour l'AFC dans le cas de deux variables, on considère un N-échantillon d'individus, répondant cette fois-ci à Q variables (ou questions), avec $Q > 2$. On peut construire de la même façon le tableau disjonctif complet et une généralisation de la table de contingence appelée *matrice du Burt.*

3.4.1 ACM sur le tableau disjonctif complet

On note $X_1, ..., X_Q$ les tableaux disjonctifs complets correspondant aux Q questions. Chaque question $q, 1 \leq q \leq Q$ possède m_q modalités et chaque individu répond à une question q en choisissant une modalité parmi les m_q modalités.
Ainsi le tableau disjonctif complet K de tout le questionnaire est la juxtaposition des tableaux $X_q, \;\; 1 \leq q \leq Q$.

$$K = [X_1 | X_2 | \ldots\ldots\ldots | X_Q] \,.$$

On note M le nombre total de modalités, et M_q le nombre de modalités des q premières questions.

$$M = \sum_{q=1}^{q=Q} m_q \quad M_q = \sum_{r=1}^{r=q} m_r .$$

Pour tout m, $1 \leq m \leq M$, on note Z^m le vecteur de $\{0,1\}^N$ composé des N réponses d'individus à la modalité m. On construit la matrice $Individus \; \times \; Variables$

$$K = (Z^1, \ldots, Z^{M_1}, \ldots, Z^m, \ldots, Z^M)$$

K	Z^1	...	Z^m	...	Z^M	K_N
1	0	...	0	...	1	Q
	1	...	0	...	0	
i	0	...	k_{im}	..	.	
		...	.	..	.	
	1	...	0	...	.	
N	0	...	0	...	1	Q
K_M	$l_{\cdot 1}$	...	$l_{\cdot m}$	...	$l_{\cdot M}$	NQ

TAB. 3.6 - *Tableau disjonctif complet*

La somme des colonnes Z^m correspondant à une même question est égale au vecteur $\mathbf{1}_N$ (vecteurs de $\mathbb{R}^N$ dont toutes les composantes sont égales à un). Le rang de la matrice K est donc égal à $M - Q + 1$.

La marginale sur les lignes est un vecteur constant égal à $Q\mathbf{1}_N$ puisqu'il y a Q questions et que pour chaque question, on ne choisit qu'une seule modalité. La transformation des lignes en profils n'a donc pas de conséquence importante dans le calcul des distances, puisque les sommes en ligne sont toutes égales.
On note L la matrice des profils donnée par :

$$L = \frac{1}{Q}K,$$

dont les éléments sont notés $l_{im}, 1 \leq i \leq N, 1 \leq m \leq M$.

La marginale sur les colonnes est le vecteur $(l_{\cdot 1}, l_{\cdot 2}, \ldots, l_{\cdot m}, \ldots, l_{\cdot M})$, où $l_{\cdot m}$ est le nombre d'individus qui ont choisi la modalité $m, 1 \leq m \leq M$.

On définit la distance du Chi-deux entre deux profils lignes l^i et l^j par

$$\chi^2(l^i, l^j) = \sum_{m=1}^{m=M} \frac{NQ}{l_{\cdot m}} (l_{im} - l_{jm})^2 = \sum_{m=1}^{m=M} \frac{QN}{l_{\cdot m}} (l_{im} - l_{jm})^2$$

On note W la matrice diagonale

$$W = Diag\,\{l_{\cdot 1}, \ldots, l_{\cdot m}, \ldots, l_{\cdot M}\},$$

alors la métrique qui est associée aux profils lignes est définie par la matrice

$$M = NQW^{-1}.$$

On suppose que les individus sont tous équiprobables et donc la matrice des poids choisie est

$$D = \frac{1}{N} I_N$$

De la même manière que pour l'AFC, les profils lignes appartiennent à un même hyperplan orthogonal au vecteur profil-ligne moyen g_l. On montre aussi que les axes principaux non triviaux de la matrice VM sont ceux de la matrice X^tDXM et que les facteurs sont vecteurs propres de la matrice MX^tDX. Donc on peut calculer ces deux matrices et on a :

$$X^tDXM = \frac{1}{Q}K^t \frac{1}{N} I_N \frac{1}{Q} KNQW^{-1} = \frac{1}{Q}K^tKW^{-1}$$

$$MX^tDX = \frac{1}{Q}W^{-1}K^tK$$

L'analyse en composantes multiples se ramène donc à une analyse en composantes principales du tableau disjonctif K, pour un choix convenable de la métrique et de la pondération. Il s'agit donc de diagonaliser les matrices K^tKW^{-1} et $W^{-1}K^tK$. Mais en fait, on s'aperçoit qu'il apparaît ici la matrice $B = K^tK$ appelée *matrice de Burt* qui est étudiée dans le paragraphe suivant.

3.4.2 ACM sur la matrice de Burt

Le tableau (ou matrice) de *Burt* peut être vue comme une généralisation d'un tableau de contingence à deux variables. En effet, c'est une matrice qui comprend les tableaux de contingence des variables prises deux à deux. Posons

$$B = K^tK$$

Alors B peut s'écrire

$$B = \begin{bmatrix} X_1^tX_1 & X_1^tX_2 & & X_1^tX_Q \\ X_2^tX_1 & & & \\ & & X_i^tX_j & \\ X_Q^tX_1 & & & X_Q^tX_Q \end{bmatrix}$$

C'est une matrice symétrique de dimension $M \times M$ qui sera aussi désignée par son terme général $(b_{ij})_{1\leq i,j\leq M}$. Le terme b_{ii} pour $1 \leq i \leq M$ est le nombre d'individus ayant choisi la modalité i, ainsi on a

$$\forall i,\ 1 \leq i \leq M,\ b_{ii} = l_{.i}.$$

D'autre part les blocs diagonaux sont des matrices diagonales, puisqu'il s'agit du croisement d'une variable avec elle même et qu'on ne choisit qu'une seule modalité dans une variable. Si on calcule à l'intérieur d'un bloc correspondant à la question q, la somme

$$\sum_{j=M_{q-1}+1}^{j=M_q} b_{ij} = l_{.i} = b_{ii},$$

on voit donc que la somme de la ligne i de la matrice B est égale à $Qb_{ii} = Ql_{.i}$.

La matrice B étant symétrique, on ne considère que les profils lignes, les autres résultats sont déduits d'une façon symétrique. La matrice des profils lignes est alors

$$F = \frac{1}{Q}W^{-1}B,$$

B	Z_1	Z_j	Z_M	b
Z_1	$\begin{matrix} b_{11} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & . \end{matrix}$			$Ql_{\cdot 1}$ . .
Z_i		b_{ij}		$Ql_{\cdot i}$
Z_M			$\begin{matrix} . & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & b_{MM} \end{matrix}$	$Ql_{\cdot M}$
b^t	$Ql_{\cdot 1}$	$Ql_{\cdot i}$	$Ql_{\cdot M}$	Q^2N

TAB. 3.7 - *Tableau de Burt*

la matrice des poids est

$$D = \frac{1}{QN}W,$$

et finalement la métrique est

$$M = QNW^{-1}.$$

On montre également que les profils lignes appartiennent à un hyperplan orthogonal au profil moyen. Ce dernier engendre l'axe principal trivial et les autres axes principaux sont engendrés par les vecteurs propres de F^tDFM. C'est-à-dire :

$$F^tDFM = \frac{1}{Q}^{2}BW^{-1}WW^{-1}BW^{-1} = \frac{1}{Q^2}BW^{-1}BW^{-1}$$

Mais si a est un axe principal de BW^{-1} avec une valeur propre λ, alors il est aussi axe principal de $BW^{-1}BW^{-1}$ avec une valeur propre λ^2.

On voit donc qu'il est équivalent de travailler à partir du tableau disjonctif K ou de la matrice de Burt B. C'est cette remarque très simple qui est à l'origine de la définition de l'algorithme KACM, qui est une adaptation de l'algorithme de Kohonen à l'étude des tables de Burt représentant les croisements entre deux ou plusieurs variables qualitatives.

Comme dans le cas de l'AFC, on peut faire une correction sur la matrice B pour se ramener à la métrique euclidienne usuelle ainsi qu'à la pondération uniforme, en effet on a :

$$\frac{1}{Q}W^{-1} = Diag\left\{\frac{1}{b_{1\cdot}}, \ldots, \frac{1}{b_{i\cdot}}, \ldots, \frac{1}{b_{M\cdot}}\right\}.$$

Si on pose

$$B^c = \frac{b_{ij}}{\sqrt{b_{i.}}\sqrt{b_{.j}}},$$

alors les axes principaux de l'ACM par le tableau de Burt sont les vecteurs propres de la matrice $(B^c)^t B^c$, c'est-à-dire que l'on se ramène à une ACP usuelle sur la matrice des données corrigées B^c.

De même, si on pose

$$K^c = \frac{k_{ij}}{\sqrt{k_{i.}}\sqrt{k_{.j}}},$$

alors les axes principaux de l'ACM par le tableau disjonctif complet sont les vecteurs propres de la matrice $(K^c)^t K^c$.

Nous présenterons dans le chapitre 4 la méthode KACP qui utilise l'algorithme de Kohonen pour réaliser une sorte de projection comparable à une analyse en composantes principales et comment à partir d'une carte de Kohonen on peut construire des super-classes. Le chapitre 6 est consacré à la méthode KACM que nous avons développée et qui permet de traiter les variables qualitatives dans le cas où l'on a deux variables ou plus.

Chapitre 4

L'algorithme de Kohonen et la représentation de données quantitatives (KACP)

4.1 L'algorithme d'auto-organisation de Kohonen

Les algorithmes neuronaux de type perceptron sont des algorithmes qui nécessitent un superviseur pour l'apprentissage du réseau. Pour la base d'apprentissage on connaît la sortie désirée correspondant à chaque « individu » . L'apprentissage étant réalisé, si une nouvelle observation se présente, le réseau est capable de la classer compte tenu de l'apprentissage préalable qu'il a subi. Ce sont des algorithmes dits « supervisés ». Cependant on peut définir des réseaux de neurones artificiels qui sont capables d'apprendre à partir de leur propre environnement. Ce sont des algorithmes dits « non-supervisés » ou « auto-organisants ». On retrouve dans cette catégorie l'algorithme d'analyse en composantes principales d'Oja et celui de Sanger, et l'algorithme d'auto-organisation de Kohonen.

Les premières travaux de Teuvo Kohonen sur les réseaux de neurones (Kohonen, 1974), (Kohonen, 1977) et (Kohonen, 1980) sont principalement consacrés autour du thème de la mémoire associative. Ses recherches ultérieures, (Kohonen, 1981), (Kohonen, 1982) et (Kohonen, 1984), l'ont ensuite mené à définir une extension de l'apprentissage compétitif appelée **Self Organizing Map** (carte auto-organisante) qui est un algorithme neuronal qui réorganise, d'une façon non-supervisée, les vecteurs codes d'un réseau de neurones en préservant la topologie de l'espace des entrées. Il définit ensuite dans (Kohonen, 1986) et (Kohonen, 1988) l'algorithme **LVQ** (Linear Vector Quantization) qui automatiquement (sans superviseur) détermine les k meilleurs vecteurs codes d'un large échantillon de points issus d'un espace vectoriel de dimension d. Il a appliqué ses algorithmes dans divers domaines tels que l'organisation des sons dans une carte phonotopique, le contrôle commande, la reconnaissance de la parole, et dernièrement la visualisation des données.

Dans le cadre de notre travail, cet algorithme est particulièrement intéressant en raison de deux caractéristiques importantes. C'est un algorithme neuronal non supervisé qui réalise d'une part une quantification vectorielle et d'autre part une classification des individus de l'espace d'entrée. En outre le résultat se représente sur une carte, d'une façon analogue à l'ACP. Le principe de l'algorithme consiste à faire l'apprentissage d'un réseau de noeuds (ou unités) disposés généralement en ligne ou en grille puis à classer les entrées dans le réseau. Les entrées sont les lignes d'un tableau de données ($n \times d$) « Individus × variables ».

4.1.1 Définitions et notations

On note $\Xi \subset \mathbb{R}^d$ l'espace des n entrées (ou stimuli) qui vont servir à l'apprentissage du réseau. Le réseau R est constitué de m unités disposées en « *grille* » ou « *ficelle* » (voir plus bas les dispositions relatives) et à chaque unité on associe un vecteur de $\mathbb{R}^d$ (de même dimension que l'espace des stimuli) appelé « vecteur poids » ou « vecteur code », on a donc

$$Y_t^1, Y_t^2, ..., Y_t^i, ..., Y_t^m$$

les m vecteurs poids à l'instant t de l'apprentissage. Ce sont m variables aléatoires d-dimensionnelles dont les réalisations à chaque instant sont notées

$$y_t^1, y_t^2, ..., y_t^i, ..., y_t^m.$$

A chaque instant, le réseau est donc défini par un état Y_t tel que $Y_t = (Y_t^1, Y_t^2, ..., Y_t^i, ..., Y_t^m)$.

D'autre part, pour introduire une notion de proximité entre unités, on définit sur les unités du réseau R une fonction de voisinage $V_{r(t)}(i)$ qui décrit les unités voisines de l'unité i; ce sont celles qui sont dans un rayon $r(t)$ autour de l'unité i. Ainsi durant l'apprentissage, les stimuli de l'espace d'entrée excitent des unités voisines au sens de cette «topologie». En pratique, cette fonction de voisinage est une fonction décroissante du temps par son rayon $r(t)$.

Dans un réseau disposé en ficelle, la figure 4.1 ci-dessous représente les fonctions de voisinage pour un rayon de 3 (7 voisins), un rayon de 2 (5 voisins) et un rayon de 1 (3 voisins).

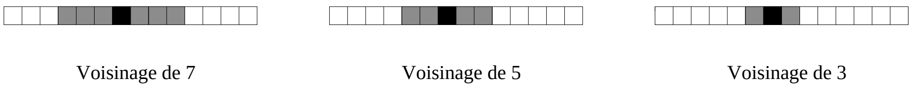


FIG. 4.1 - *Voisinage pour un réseau ficelle.*

De même, dans le cas d'un réseau grille, la figure 4.2 représente les voisinages pour les mêmes rayons, ce qui donne 49, 25 et 9 voisins.

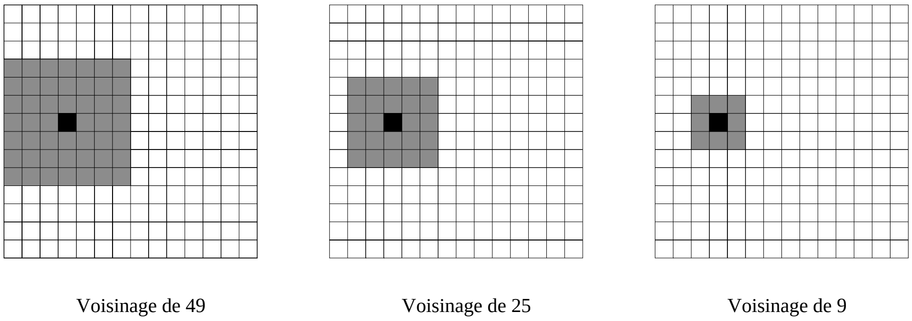


FIG. 4.2 - *Voisinage pour un réseau grille.*

On commence d'abord par initialiser aléatoirement les m vecteurs poids du réseau. Le principe de l'algorithme peut être décrit en trois étapes :
-1) on présente au réseau un individu choisi selon une distribution de probabilité $P = (P_1, P_2, ..., P_n)$ qui pondère les éléments de l'ensemble Ξ. En pratique on choisit la loi uniforme $(\forall j, \quad 1 \leq j \leq n, \quad P_j = \frac{1}{n})$, ou dans certains cas on se ramène à cette loi uniforme via une transformation des données.
-2) on recherche l'unité gagnante parmi toutes les unités du réseau. C'est celle qui réalise le minimum de la distance euclidienne entre l'individu et les vecteurs poids des unités du réseau.
-3) on réactualise les vecteurs poids de l'unité gagnante et de ses voisines. Ces trois étapes sont réitérées jusqu'à convergence pour obtenir un « état organisé » ou encore « ordonné » du réseau.

L'algorithme

I On initialise aléatoirement les vecteurs poids du réseau.

II A l'instant t on présente une entrée choisie parmi la base de données d'apprentissage, c'est-à-dire un vecteur $\omega(t+1)$ de $\Xi \subset \mathbb{R}^d$.

III On recherche l'unité gagnante i_0, définie par :

$$i_0 = \arg \min_{1 \leq i \leq m} \|\omega(t+1) - y_t^i\|^2$$

IV On modifie ensuite la valeur du vecteur poids pour l'unité gagnante et ses voisines par,

$$\begin{cases} y_{t+1}^i = y_t^i + \epsilon(t)\,(\omega(t+1) - y_t^i) & \forall i \in V_{r(t)}(i_0) \\ y_{t+1}^i = y_t^i & \forall i \notin V_{r(t)}(i_0) \end{cases}$$

où $\epsilon(t)$ est un paramètre d'adaptation qui vérifie les conditions de l'algorithme de Robins-Monroe :

$$\sum_t \epsilon(t) = \infty, \quad \sum_t \epsilon(t)^2 < \infty.$$

La fonction de voisinage dépend du temps par son rayon $r(t)$, et pratiquement on commence avec un rayon assez grand (quatre ou cinq ou plus, tout dépend de la taille du réseau) et on fait décroître ce rayon jusqu'à la valeur zéro, ce qui correspond alors au cas à zéro-voisin qui n'est autre que l'algorithme k-means (MacQueen, 1967).

Après apprentissage, chaque individu est classé dans l'unité du réseau qui lui est la plus proche au sens de la distance quadratique. Or chaque unité i du réseau est caractérisée par un vecteur code y_T^i. Ceci définit donc un quantificateur vectoriel q de la manière suivante : si un individu x_j est classé dans l'unité i on a,

$$x_j \quad \xrightarrow{q} \quad q(x_j) = y_T^i,$$

où y_T^i est la valeur finale du vecteur poids de l'unité i. On définit comme d'habitude les classes C_i par

$$\forall i,\ 1 \leq i \leq m, \quad C_i = q^{-1}(y_T^i) = \{x_j \in \Xi, \quad \text{qui sont classés dans l'unité } i\},$$

Ce sont les m classes générées par l'algorithme. Il y a au plus m classes, car il se peut que certaines de ces classes soient vides. Alors

$$\mathcal{C} = \{C_1, ..., C_m\}.$$

est la partition de Voronoï de l'ensemble Ξ. Le vecteur code de la classe C_i est le vecteur poids final y_T^i de l'unité i. En fait la structure de voisinage définie sur les unités, permet de définir une structure de voisinages sur les classes. On obtient ainsi une carte où sont projetés les individus et qui respecte la topologie de l'espace d'entrée des individus. En d'autres termes deux individus qui sont voisins au sens de la distance dans $\mathbb{R}^d$ sont classés dans une même unité ou dans deux unités voisines.

Un des points qui rendent difficile l'étude théorique dans le cas général de l'algorithme de Kohonen, est qu'il n'existe pas de potentiel (ou énergie) associée lorsque les entrées sont munies d'une loi de probabilité P continue (voir dans (Erwin et al., 1992a) et (Erwin et al., 1992b)). Mais au contraire, dans le cas qui nous intéresse ici, c'est-à-dire quand l'espace des entrées est fini, muni d'une probabilité discrète, (Ritter et al., 1992) ont montré que l'algorithme de Kohonen pour un rayon de voisinage r constant, est un algorithme du gradient stochastique qui minimise le potentiel :

$$V(Y) = \sum_{i=1}^{m} \sum_{k \in V_r(i)} \sum_{x_j \in C_k} \|x_j - y^i\|^2,$$

où $Y = (y^1, ..., y^m)$.

Ce potentiel généralise la variance intra-classes. Ici on calcule la somme des carrés des distances de chaque observation non seulement à son vecteur code mais aussi aux vecteurs codes des classes voisines.

La distorsion qui est associée à cette quantification est :

$$D_q = \sum_{i=1}^{i=m} \sum_{x_j \in C_i} \|x_j - y_T^i\|,$$

où seules les distances aux centres des classes sont prises en compte.

Nous constaterons sur des exemples, que cette distorsion fluctue autour d'une valeur moyenne quand on répète plusieurs fois l'algorithme. Cela montre que malgré les petites différences de classement qu'on constatera entre deux versions de l'algorithme, globalement l'organisation finale reste la même.

4.2 Une méthode analogue à l'Analyse en Composantes Principales

Dans le cas de l'analyse des données, on utilise l'algorithme de Kohonen pour « projeter » les individus d'une matrice X $individus \times variables$ sur une carte de Kohonen. Ce n'est pas une projection de type ACP (projection orthogonale sur des plans successifs), mais plutôt une classification des individus dans des unités (ou classes) qui sont disposées en respectant la fonction de voisinage. Le résultat de l'algorithme est une carte qui ressemble fortement à celle qu'on obtient à l'aide d'une ACP. Evidemment on n'en obtient qu'une, qui est forcément assez grossière en comparaison des projections planes successives obtenues en ACP classique.

Cependant une carte de Kohonen a certains avantages par rapport à une projection linéaire type ACP. Premièrement, celui de résumer plusieurs projections en une seule tout en préservant le mieux possible la topologie de l'espace d'entrée, alors que l'ACP est une technique factorielle linéaire qui projette le nuage des individus sur des axes maximisant l'inertie projetée et qui nécessite par conséquent la consultation de plusieurs projections successives si on veut éviter des erreurs grossières. Deuxièmement l'algorithme de Kohonen produit une classification et une quantification, puisque chaque individu est classé dans une unité et représenté par le vecteur poids correspondant.

Le choix du nombre d'unités dépend principalement du nombre d'individus et il n'y a pas d'argument théorique qui permette de le calculer numériquement, mais seulement des propositions empiriques. D'un autre côté, la géométrie du réseau dépend de la forme du nuage de points : si celui-ci est plutôt allongé, il vaut mieux utiliser un réseau rectangulaire dont la longueur sera plus importante que la largeur; si les variables sont centrées réduites, on choisit en général un réseau carré; si une direction prédomine, un réseau en ligne sera très utile etc...

Au sujet du temps d'apprentissage T, on estime qu'il faut faire passer en moyenne cinq fois chaque individu de l'ensemble Ξ. Si l'ensemble Ξ contient un nombre très grand d'individus, on peut en extraire un sous-ensemble représentatif qui sert pour l'apprentissage du réseau. Après l'apprentissage, on classe tous les éléments de Ξ dans le réseau.

A la fin de l'apprentissage, on obtient une carte ayant au plus m classes, et on peut diminuer le nombre de classes en regroupant les classes de Kohonen en super-classes. Pour cela on peut faire une classification hiérarchique ou utiliser l'algorithme de fusion 2.7.4 sur les vecteurs poids des unités. Les super-classes ainsi obtenues sont généralement bien connexes et contiguës. Ce qui confirme la convergence de l'algorithme. Tout ceci est illustré dans les paragraphes suivants.

4.3 Exemples

4.3.1 Des exemples dans l'espace $\mathbb{R}^2$

Les premiers exemples que nous présentons dans ce paragraphe sont tous de dimension deux, c'est-à-dire que l'espace d'entrée Ξ est un sous ensemble de $\mathbb{R}^2$. Ces exemples montrent comment le réseau de Kohonen reconstitue la topologie de l'espace des stimuli. Comme on est dans le plan on peut donc représenter les positions finales des différentes unités du réseau dans les cas d'une grille et d'une ficelle. Dans le premier exemple illustré par les figures 4.3 et 4.4, on fait l'apprentissage d'un réseau de Kohonen en grille puis en ficelle à partir de 5000 points tirés suivant une loi uniforme sur un intervalle carré de longueur 10. Comme la dimension est deux, on peut représenter les positions finales des vecteurs poids des différents unités du réseau. On a représenté, dans les figures ci-dessous, simultanément les points de la base d'apprentissage et la disposition finale dans l'espace des vecteurs poids des unités du réseau.

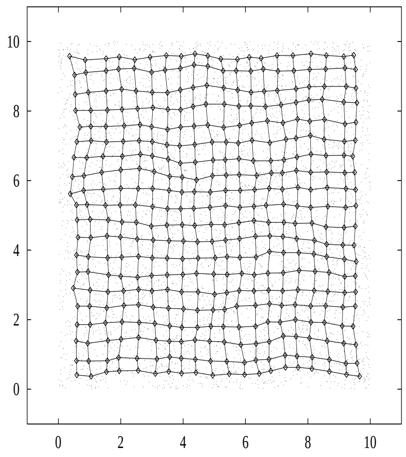


FIG. 4.3 - *Une grille; Ξ est formé de 5000 points dans un carré.*

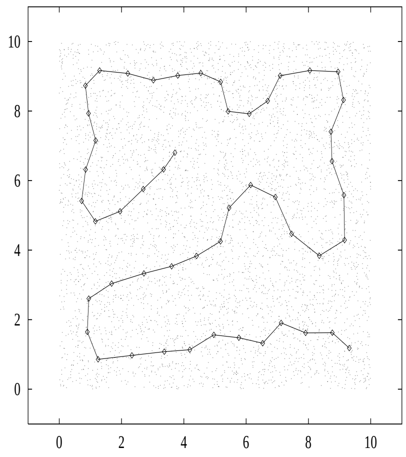


FIG. 4.4 - *Une ficelle; Ξ est formé de 5000 points dans un carré.*

Deux remarques peuvent être faites sur ce premier exemple : d'une part, dans les deux cas, (grille et ficelle), l'algorithme fait une très bonne quantification de l'espace Ξ; d'autre part à cause de la notion de voisinage intrinsèque au réseau on obtient, dans le cas de la ficelle, une sorte de projection curviligne qui est beaucoup moins pénalisante que celle obtenue par l'ACP, puisque en ne gardant qu'un axe principal, on obtiendrait l'un des deux axes diagonaux et la projection serait très déformée.

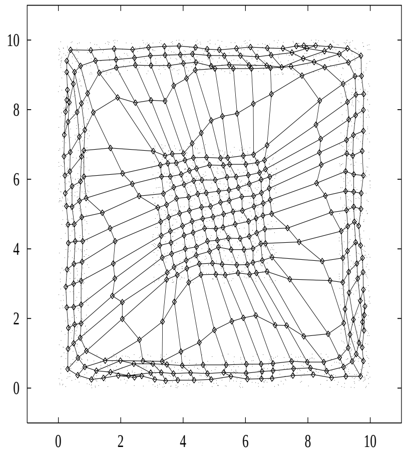


FIG. 4.5 - *Une grille; Ξ est formé d'un petit carré au centre et d'un cadre.*

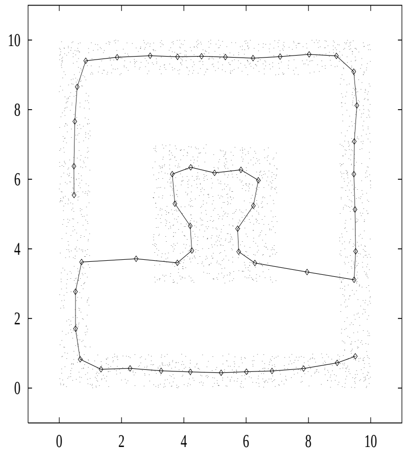


FIG. 4.6 - *Une ficelle; Ξ est formé d'un petit carré au centre et d'un cadre.*

Dans l'exemple suivant (fig 4.5), on reprend l'exemple précédent en enlevant une couronne de points. On remarque alors que l'algorithme de Kohonen essaye bien de reproduire cette topologie en refoulant vers le centre et les bords les unités qui étaient auparavant situées dans cette région. Le réseau ficelle (fig 4.6) déplie l'espace en le découpant au niveau des deux bords de la ficelle.

On peut faire, sur ce deuxième exemple, une remarque très utile pour la lecture des cartes de Kohonen. En effet ce que l'on constate sur la figure 4.5, et d'une façon plus évidente que sur la figure 4.6, c'est que les unités qui restent entre la couronne et le carré central vont être, pour la plupart, des unités vides car il n y a pas de stimuli proches. On en déduit donc, sans que cela soit une règle, que les unités vides sur une carte de Kohonen présupposent l'existence d'une frontière entre de groupes d'unités.

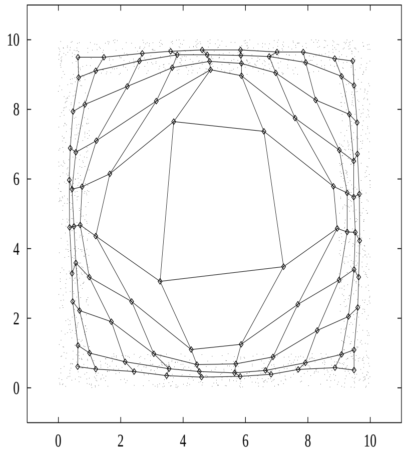


FIG. 4.7 - *Une grille*

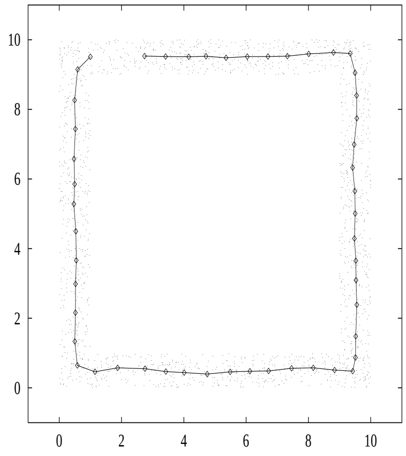


FIG. 4.8 - *Une ficelle*

Pour le troisième exemple ci-dessus, on ne garde que la couronne extérieure. La grille (fig 4.7) se concentre sur les bords et la ficelle (fig 4.8) déplie cette couronne en la découpant au niveau de ses deux bords. On peut dire que la ficelle réalise une très bonne quantification.

Ceci étant, les deux unités extrêmes représentent deux régions voisines du plan et pour avoir une quantification optimale, il faudrait décider que ces deux unités sont elles aussi voisines.
Il arrive parfois que l'algorithme de Kohonen converge vers des états non ordonnés. Les figures 4.9 illustrent un exemple d'une telle situation qui se produit dans $\mathbb{R}^2$ et qu'on appelle un « *papillon de Kohonen* ». On peut se rendre compte de ces situations, soit visuellement en dimension deux, soit parce qu'on obtiendra des super-classes mal constituées, quand on regroupe les classes de Kohonen en super-classes au moyen d'une classification hiérarchique par exemple.

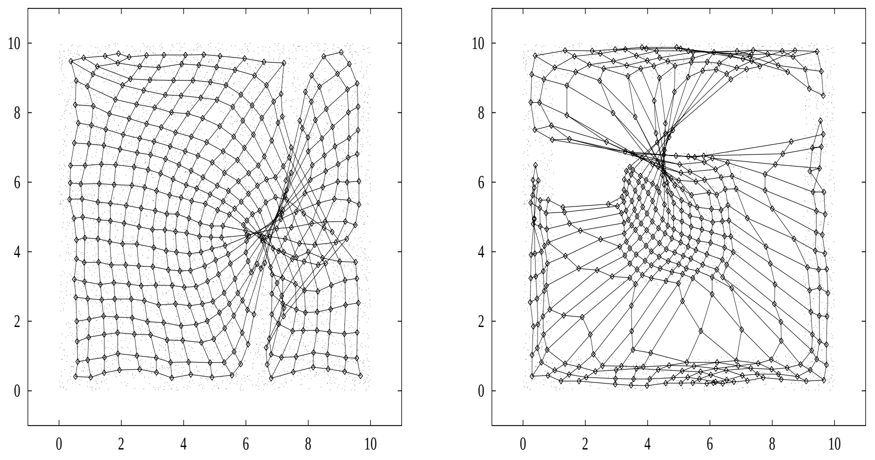


FIG. 4.9 - *Papillons de Kohonen*

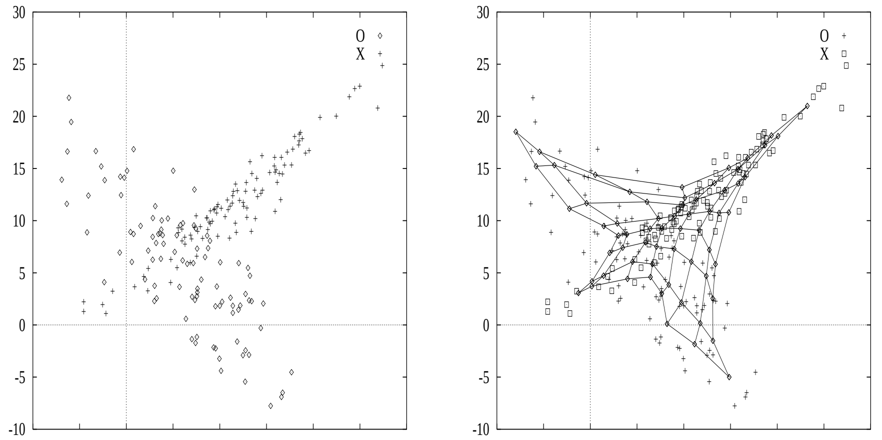


FIG. 4.10 - *Deux échantillons gaussiens*

Dans l'exemple suivant, il s'agit de deux échantillons gaussiens (symbolisés l'un par «X» et l'autre par «O») qui s'entrecoupent. C'est un cas de recouvrement de classes qui peut être

traité par des méthodes linéaires. On voit dans la figure 4.10 comment le réseau épouse la forme des deux nuages.
On a un scénario presque identique dans l'exemple suivant illustré par la figure 4.11 avec trois classes qui se recouvrent. Il faut cependant noter que les noeuds qui se trouvent sur les coins de la carte ne vont pas jusqu'aux bords de l'échantillon. C'est dû au fait que celui-ci est plus dense au centre, on tire durant l'apprentissage plus de points dans la région du centre.

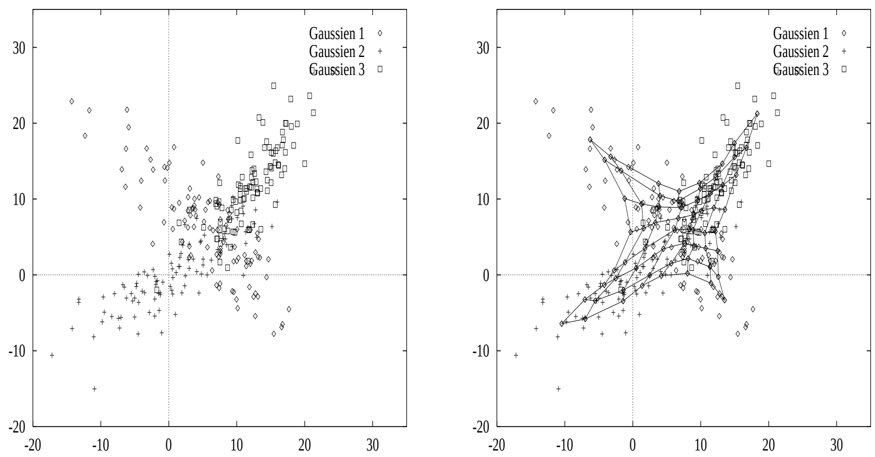


FIG. 4.11 - *Trois échantillons gaussiens*

Les figures 4.12 montrent un ensemble Ξ formé de 3000 points choisis uniformément dans un cercle de rayon 5, le réseau se déforme alors pour prendre la forme d'un cercle.

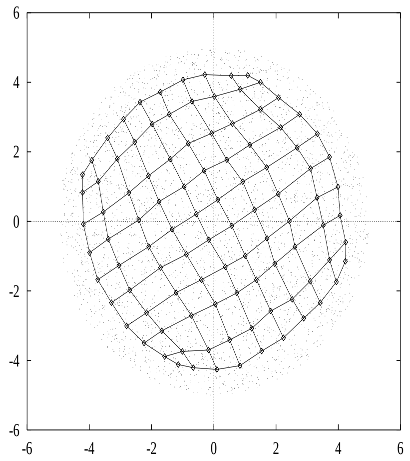


FIG. 4.12 - *Apprentissage d'un cercle*

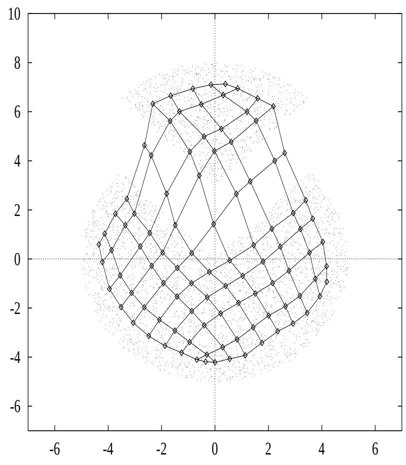


FIG. 4.13 - *Un cas non linéairement séparable.*

La figure 4.13 montre un cas non linéairement séparable et pour lequel l'algorithme de Kohonen donne un très bon résultat. En effet, on remarque la régularité les mailles à l'intérieur de chacune des classes, par contre entre les deux classes les mailles sont beaucoup plus grandes.

Ces quelques exemples illustrent la façon dont l'algorithme conserve la topologie de l'espace de départ. Nous allons à présent examiner un exemple de traitement de données économiques réelles.

4.3.2 Une représentation du développement économique

Cet exemple pris de (Letremy, 1997) suivant (Blayo and Demartines, 1991) représente les valeurs, en 1996 et dans 96 pays, de sept variables socio-économiques qui sont :
le taux de croissance annuelle, le taux de mortalité infantile,
le taux d'analphabétisme, le pourcentage de scolarisés dans le second degré,
le PNB par habitant, le pourcentage de chômeurs dans la population active et
le taux d'inflation.

Généralement et comme cela se pratique en ACP, lorsqu'il s'agit de données quantitatives continues, on centre et on réduit les variables pour éviter l'effet « taille ». On a donc une matrice X constituée de sept variables centrées réduites en colonnes et de 96 lignes qui représentent les pays.

	0	1	2	3	4	5	6	7
0	Albanie Guyana Vietnam C	Paraguay	Laos	Arab-Sao Nicaragua Syrie	Comores	Cote d'Ivoire Ghana Maroc Mauritanie Pakistan	Soudan Yemen	Angola Afghanistan D
1	Venezuela	Salvador	Bolivie	Swaziland		Cameroun	Nigeria	Haiti Mozambique
2	Costa-Rica Thailand	Equateur Mexique	Turquie	Mongolie Perou	Egypte Tunisie	Algerie		Indonesie Kenya Namibie Zimbabwe
3	Malaisie	Colombie Panama	Philippines	***********	***********	***********	Afri-sud Iran Liban	Macedoine
4	Chili Chine Fidji		Sri-Lanka	Jamaique	Argentine	Bahrein	***********	***********
5	Moldavie	Uruguay	Bulgarie Pologne Yougosla			Singapour	Emi-arab Israel	Australie Canada
6	Ukraine	Russie	Croatie Hongrie	Slovenie	Portugal	Italie Nou-Zelande Roy-Uni	Suede	Etats-Unis Islande
7	Brésil B		Roumanie	Tchequie	Chypre Coree-Sud Grece Malte	Espagne Finlande Irlande	Allemagne Belgique Danemark France Norvege Pays-Bas	Luxembourg Japon Suisse A

TAB. 4.1 - *Carte de Kohonen*

Avec les lignes de cette matrice X, on procède à l'apprentissage d'une grille de Kohonen constituée de 8×8 noeuds. Après 1000 itérations (soit l'équivalent de 3 secondes de calcul sur une station Sun Sparc 5), on obtient la grille représentée dans le tableau 4.1. Afin de mieux faciliter la lecture de la carte, on numérote les unités par un couple d'entiers tel que le premier entier numérote les unités de haut en bas et le second de gauche à droite. Ainsi le couple (0,0) correspond à la première unité en haut à gauche et le couple (7,7) à la dernière unité en bas à droite.

En observant attentivement la carte de Kohonen (Tab 4.1), on constate que ses quatre coins ont été attribués d'une façon plus ou moins hiérarchique. En effet, les pays « riches» tels que les Etats-unis, la France, l'Allemagne... se regroupent dans les unités en bas à droite autour de l'unité (7,7), dans une région qu'on désigne par A. On se déplace ensuite vers la zone B qui se situe en bas à gauche autour de l'unité (7,0) où l'on retrouve beaucoup de pays de l'Europe de l'Est (Russie, Ukraine, Pologne...) ainsi que des pays semi-industriels comme le Brésil et l'Uruguay. Puis on remontant vers le coin en haut à gauche autour de l'unité (0,0), qu'on appellera la zone C, on retrouve certains pays d'Asie (Chine, Malaisie, Vietnam) et d'autres d'Amérique du sud (Paraguay, Vénezuela, Bolivie...). Enfin en finissant la boucle, on arrive à la zone D située en haut à droite de la grille où sont classés les pays « pauvres » qui sont principalement des pays d'Afrique. Remarquons la frontière entre la zone A et la zone D constituée d'unités vides et qu'on a signalé par des astérisques.

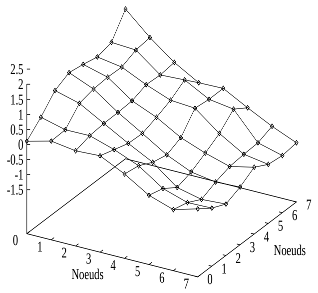


FIG. 4.14 - *Croissance annuelle*

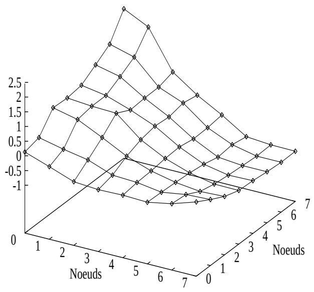


FIG. 4.15 - *Mortalité infantile*

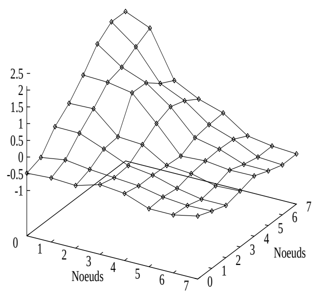


FIG. 4.16 - *Analphabétisme*

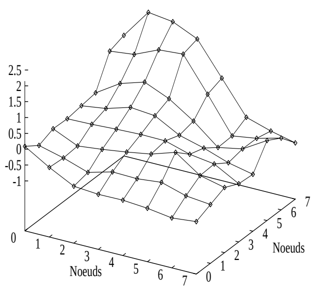


FIG. 4.17 - *Chômage*

D'autre part, on peut regarder les profils des variables à travers les unités de la grille. Les figures 4.14 à 4.20 représentent les valeurs des composantes des vecteurs poids pour chacune des unités de la grille. Comme les valeurs des stimuli sont centrées et réduites, on a donc des valeurs de poids négatives. On ne s'intéresse pas ici aux valeurs absolues mais à la comparaison des niveaux relatifs entre les différentes zones A, B, C et D.

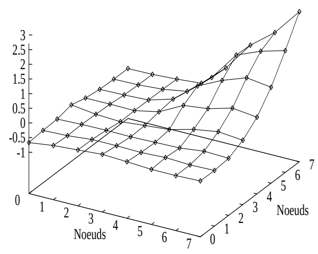


FIG. 4.18 - *PNB*

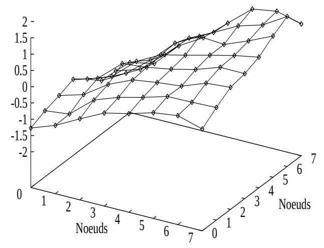


FIG. 4.19 - *Etudes Secondaires*

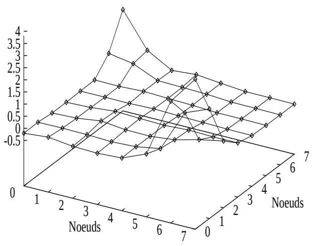


FIG. 4.20 - *Inflation*

Ainsi la figure 4.14 illustre le comportement de la variable « croissance annuelle » qui révèle un axe diagonal depuis la zone B (faible croissance) vers la zone D (forte croissance). De même les indicateurs, « taux d'analphabétisme », « taux de mortalité infantile » et « taux de chômage », figures 4.15, 4.16 et 4.17, ont des variations très semblables, avec le minimum en A et B et le maximum en D. Pour la variable « Etudes secondaires », on a le schéma inverse puisque le maximum est en A, puis la variable décroît vers B puis vers C et D. Le PNB est de son côté pratiquement constant dans les zones B, C et D et très élevé dans A. L'inflation présente deux pics dans B et D, ailleurs et relativement à ces pics, on peut dire quelle est pratiquement uniforme. On peut citer le Brésil qui en ayant la plus forte inflation (1024), se retrouve seul dans la classe (7.0). L'Ukraine et la Russie, ses deux plus proches voisins, sont eux aussi ceux qui enregistrent le plus fort taux d'inflation. Ceci permet l'interprétation des classes et des zones de la carte.

4.3.3 Etude du marché immobilier francilien

Il s'agit d'une étude pour le compte de la Direction de l'habitat et de la ville de la Direction Régionale de l'Equipement d'Ile de France (DHV/DREIF), et qui a été réalisée en collaboration avec le laboratoire d'économie urbaine (C3E/METIS)[1] de l'université Paris I. Il s'agissait d'étudier les relations quantitatives entre valeurs immobilières et foncières en Ile de France, sur les marchés des logements et des bureaux. Un article reprenant les résultats de cette étude a été publié dans la revue «*International journal of urban and regional research* » (Gaubert et al., 1996). Il est donné en annexe. On se contente donc ici de résumer les principaux résultats.

L'étude repose sur l'exploitation d'une base de données constituée de la réunion des fichiers IMO de la DGI (Ministère des finances) et SICLONE[2] de la DAEI (Ministère de l'équipement), auquels il faut rajouter des données OLAP[3] sur les loyers du secteur privé et HLM sur les loyers du secteur social. D'autre part, des indicateurs socio-économiques sont tirés du recensement général de la population (RGP-INSEE). En définitive, on dispose d'une série d'observations sur 16 ans (1975-1991) et pour 256 communes d'Ile de France qui se répartissent comme suit :

Paris:	20
Hauts-de-seine :	38
Val de Marne :	38
Essonne :	37
Seine Saint-Denis :	25
Yvelines :	41
Seine et Marne :	22
Val d'Oise :	35

L'un des objectifs était d'observer comment les mécanismes de formation des prix peuvent se déformer dans le temps. Compte tenu de l'importance du retournement de conjoncture intervenu au cours des années quatre-vingt, on a divisé l'intervalle de temps en deux sous-périodes[4] :

- T_1, de 1976 à 1983 inclus;
- T_2, de 1984 à 1991

Ce découpage est fondé sur une observation des conjonctures (graphiques à prix constants), et a été validé par des analyses de données.

La recherche d'un découpage du marché francilien en segments de marchés différenciés répondait de son côté à une double préoccupation. De façon très générale, la segmentation

1. Centre d'Economie de l'Environnement et de l'Espace
2. Système d'Information sur la Construction de LOgements NEufs.
3. Observatoire des Loyers de L'Agglomération Parisienne.
4. Les observations s'arrêtant en 1991, il n'y avait pas lieu de distinguer une troisième période, correspondant au retournement du marché après 1990, et qui eut un nombre d'observations insuffisant pour établir des modèles significatifs

des marchés du logement fait partie, depuis la grande vague d'urbanisation suscitée par la révolution industrielle, des caractéristiques structurelles des villes modernes, et trouve sa traduction dans une fragmentation de l'espace résidentiel, le marché immobilier servant ainsi de vecteur, ou de filtre, à un processus de ségrégation sociale.

Par ailleurs, l'objet de l'étude C3E/DREIF était de mesurer le degré de corrélation (ou de dépendance) des prix du sol avec les prix de vente (ou de location) de la construction neuve et des logements anciens, la question-guide étant de savoir dans quelle mesure le niveau actuel des valeurs foncières en Ile de France autorise encore la production de logements neufs dans les fourchettes de loyers et/ou de prix de vente compatibles avec la solvabilité de la demande. Cette question se pose depuis longtemps pour Paris et une grande partie de la petite couronne, et pour le logement social dans son ensemble. Mais compte tenu de la forte hausse des prix du sol enregistrée dans la seconde moitié des années quatre-vingt, elle se pose aujourd'hui pour la quasi totalité de l'agglomération, et pour la construction privée de logements dits « intermédiaires ». Son acuité est cependant variable selon le type de logement, et selon sa localisation.

Dans cette perspective, il était intéressant de décomposer le marché immobilier francilien en plusieurs secteurs, ou segments, ayant à la fois un contenu géographique, et une signification socio-économique, en termes de type de peuplement, de qualité du parc immobilier et de niveaux de prix. C'est pourquoi le premier travail a consisté à dégager un classement hiérarchique des communes d'Ile de France. Sur le base de ce classement on a ensuite cherché à identifier des modèles structurels de détermination des prix fonciers et immobiliers, différenciés dans le temps, et en fonction de l'appartenance à tel ou tel segment du marché.

Compte tenu des objectifs de l'étude, on s'est limité à une partition de l'Ile de France en 3 classes. Ce choix d'une tripartition, plutôt que d'un découpage en 5, 10 ... ou n classes, est le résultat d'un compromis entre deux types de considérations :
- d'une part, la nécessité de limiter le nombre de classes, pour ne pas aboutir à des effectifs trop faibles, ce qui aurait à la fois limité la validité des modèles économétriques et compliqué leur interprétation;
-d'autre part, la nécessité, pour ne pas simplifier outrageusement la description des marchés, d'introduire plus de 2 classes. En effet, ni la structure sociale de la société française, ni l'organisation spatiale des villes ne se réduisent à une opposition binaire entre riches et pauvres, et entre beaux quartiers et zones défavorisées. L'introduction dans l'analyse d'un segment «intermédiaire», outre qu'elle est susceptible de traduire la présence de classes moyennes, est aussi le moyen privilégié de faire apparaître des phénomènes de report de déséquilibres, et de diffusion des mouvements de prix, ce segment pouvant être le lieu privilégié de transmission des tensions apparues à l'un ou l'autre extrême du marché.

Nos trois classes ayant un contenu spatial, elles ont un sens, non seulement en termes d'appartenances socio-professionnelles (et donc de pouvoir d'achat immobilier), et de niveau de prix, mais également en termes de qualité de l'espace (profil du parc immobilier, degré de centralité, etc). Elles devraient donc être pertinentes pour une étude des comportements de localisation. Le segment «supérieur» a vocation à recevoir les localisations résidentielles «libres», c'est-à-dire celles des ménages disposant d'un pouvoir d'achat immobilier suffisant pour choisir de façon plus ou moins discrétionnaire leur lieu de résidence, tandis que le segment «ordinaire» est le refuge des localisations «contraintes», c'est-à-dire dominées par des considérations de minimisation du coût du logement, et faisant pas ou peu appel à de véritables choix de localisation de ménages. Cela correspond aux zones dans les quelles une fraction significative de l'offre est le fait du secteur social. Quant au segment intermédiaire, il devrait être le réceptacle des effets d'éviction (du secteur supérieur) et de répulsion (vis à vis du secteur ordinaire) subis par les ménages et les constructeurs.

Le segment supérieur (51 communes) a la forme d'un ruban étalé sur un axe Est-Ouest, de St Maur à St Germain en Laye, doté d'une excroissance sud-ouest, et auquel s'ajoutent quelques enclaves au Nord et au Sud. Le segment ordinaire (81 communes) est assez concentré au Nord (Val d'Oise et Seine Saint-Denis) et au Sud-Est (Val de Marne et Essonne). Le segment intermédiaire (123 communes) est le plus également réparti entre les départements, avec une concentration particulière au Sud de l'agglomération. Alors que Paris se range presque entièrement dans le segment «supérieur» (seuls les 18^{eme}, 19^{eme} et le 20^{eme} sont affectés au segment du logement intermédiaire), les trois quarts de la Seine Saint-Denis sont affectés au segment «ordinaire», tandis que le segment «supérieur» y est presque vide, de même qu'en Seine et Marne et dans l'Essonne. Le tableau suivant reproduit la répartition des trois segments à travers les huit départements pour la période 1984 à 1992.

Département	Supérieur	Intermédiaire	Ordinaire
Paris 75	17	3	0
Petite Couronne			
Val de Marne 94	8	23	8
Hauts de Seine 92	16	16	3
Seine Saint-Denis 93	1	5	19
Grande Couronne			
Essonne 91	1	24	12
Seine et Marne 77	1	9	12
Yvelines 78	7	22	12
Val d'Oise 95	5	16	14

TAB. 4.2 - *Répartition hiérarchique en trois catégories en 1990*

4.3.4 Marché du logement et marché du travail en Ile de France

Il s'agit d'une étude commandée par la DATAR et le Plan Urbain aux laboratoires METIS et SAMOS[5] et qui a pour objectif de caractériser la division économique et sociale de l'agglomération parisienne, d'en établir les formes et d'en appréhender les mutations depuis une vingtaine d'années. Pour cela, on nous demande de construire une interface entre l'espace du logement et l'espace du travail. Elle se manifestera par un découpage résultant de la confrontation entre les deux dimensions, et d'établir ensuite les relations entre variables logement et variables travail propres à chaque zone identifiée au préalable. L'étude comprend deux volets, le premier empirique et statistique et le second interprétatif des conjonctures et des dynamiques économiques. Nous insisterons donc surtout sur la méthodologie utilisée et quelques résultats de la partie empirique.

La base de données est constituée de données représentatives de la position relative de la commune dans l'espace francilien en termes de variables :

i) immobilières (prix, transactions, construction neuve, qualité du parc)
ii) peuplement (structure socio professionnelle et démographique, etc)
iii) emploi et activité au lieu de résidence
iv) emploi et activité au lieu de travail.

Les trois années d'études (étude des mutations) sont celles des trois derniers recensements, soit les RGP[6] 1975, 1982 et 1990. C'est le cas des variables socio-économiques, pour les variables immobilières, elles proviennent de la base de données de l'étude DREIF (cf.4.3.3) qui est annuelle, on a par conséquent calculé dans ce cas des moyennes autour de ces trois dates. Ce sont des variables telles que les prix du sol, des logements et des bureaux, les loyers des logements et des bureaux et d'autres sur la qualité du parc telles que l'ancienté, le confort et la taille des logements (la part des grands logements (+5 pièces) et celle des petits logements). Les variables socio-économiques sont issues des RGP : part des jeunes dans la population, part des locataires, part de la population occupant un logement sans confort, taux d'actifs dans la population, taux de chômage, part des cadres dans la population active, part d'artisans et commerçants et autres.

En ce qui concerne les communes, on a dans la base de données 285 communes de plus de 10000 habitants sauf pour 27 d'entre elles qui ont moins de 10000 habitants mais qui sont dans l'agglomération. Dans ces 285 communes, chaque arrondissement de Paris est compté comme une commune. Les premiers traitements statistiques ont été réalisés sur l'année 1990 et 1982 sans les arrondissements de Paris avec un total de 266 communes.

On dispose à l'origine d'une table de 92 variables pour l'année 1982 et de 128 variables en 1990. Ces nombres sont importants et il y a eu une première sélection de variables à partir des tableaux de corrélations. On a retiré les variables redondantes en privilégiant celles ayant le moins de valeurs manquantes, ce qui nous a ramené à une base contenant 45 variables pour 1990 et 36 variables pour 1982. En consultant les résultats des ACP, on a retiré d'autres variables qui ne contribuaient que faiblement aux six premiers axes

5. Statistique Appliquée et MOdélisation Stochastique
6. Recensement Général de la Population

principaux, ce qui nous a amené à une moyenne de 20 variables pour les deux années (qui ne sont pas forcément les mêmes aux deux dates).

Nous avons réalisé des ACP pour les deux dates puis des cartes de Kohonen avec 10×10 unités. Ces cartes de Kohonen nous donnent un peu moins d'une centaine de classes environ qu'on regroupe ensuite en super classes par le biais d'une classification hiérarchique sur les vecteurs codes. La carte de Kohonen représentée dans le tableau 4.3 est le résultat obtenu pour une table 1982. Sur cette carte, les communes sont identifiées par un nom en quatres lettres (voir l'Annexe B pour le tableau des noms complets correspondants).

	0	1	2	3	4	5	6	7	8	9
0	1 MRGE NEUS SMAN SMOR	1 BOUL CLIC ISSY	1 ASNI PUTO	1 AUBV PREG SDEN SOUE	1 ROIF		4 ETAM LIVR NEUP VROI	4 VIRO PAVB RAIN DEUI GROS	4 HOUD VESI	4 PORC SACL ORME VAUR
1	1 LEVA VINC		1 BAGT LILA PANT CHTN IVRY KREM	1 SURE ALFO ACRU GENT	1 MALA CHOI	4 BOUR DRAN PIER	4 ARPA VMOM JOIN	4 PROV VILM	4 BRIE VILP MAGN	4 GARV PERY
2	2 CBVE	2 COLO MTRE SMAU	3 BOIC GARE NOGM PERX	3 SCLO VANV ENGH	3 FONT CHON CACH	3 JUVI CLAM	5 CHAT PECQ	5 CHES LIMA SART DOUR VPIN	4 MOIS STFA	4 ANDR ESSA MAUR
3	2 SGER VERS SQYa		3 MAIL FONB	3 MALF	3 BGLR CHAV FONR GARC MEUD SCAU SEVR VAVR		5 MELU VIRY		8 ELAN	
4	2 HOUI CERG	2 RUEI	3 ANTO	3 BGNX	3 CHMA RONY CHEV HAYR	5 AVON CRET	5 MEAU ULIS	5 EVRY	8 GUYA	8 CHSM LOGN NOIL SAVI TORC COIG MONB ERAG JOUY
5	2 MANJ POIS NEUM	2 CHPY VITR ARGE	3 NANT PLES	5 FRES THIA VJUI	5 RUNG	5 CHEN	5 SVRA BOIL PTOI SANN	5 MEES NOIG		8 VOIS
6	6 AULN	6 MURO ATHI ERMO	6 VELI MASS SARC VBEL	5 LONG RISO	5 CHIL EPIS GRIG FRAN	5 PLET	7 VAIR BRYM DOMO BRIC	7 RAMB	7 MCOR TAVE	8 COUR
7	6 MONT GENN BOBI LACO NOIS STAI	6 FONF SCYR EPNS	6 CELL	5 DAML SMIC GRAT OUEN	5 CLAY PLAI CLIB LIME	5 MARL	7 CHEL YERE VAUC MMOR SOIS	7 CONF VERB MFER	7 MSON HERT SLEU	7 COMB PONT ROIB GIFS
8	6 VERI VGAR ORLY VALE		6 AUBG VIGS GONE	7 BRET	7 VERN MMAG	7 LAGM MORS	7 BRUN CROS SUCY EAUB	7 ORSA SAVO ARNO CORM ISLE		7 DAMG LESI BURE
9	6 CHAN TRAP ISTD VTAN GARG	6 ACHE CARP BONE	6 CARS CORB BLAN BOND ROMA VSTG BEZO	7 COUL PERS	7 NEMO DRAV MONG GOUS	7 SGEN TREF	7 BOIA MANV IGNY PALA GAGN	7 AUFF VIPE MENN MORA MARN OSNY	7 OZOI	7 MIMO STPI EPIO AUVE BEAU

TAB. 4.3 - *Une Carte de Kohonen segmentée en huit classes par une CH*

Les « secteurs urbains », obtenus en considérant à la fois les prix, l'activité immobilière et la « position urbaine », sont sensiblement différents des segments de marché issus de l'étude DREIF/METIS de 1994 (Cf 4.3.3). On le constate sur la carte de Kohonen ci-dessus qui correspond à l'année 1982. Les super-classes sont très hétérogènes du point de vue des prix. On peut cependant les ordonner à partir d'un double critère de niveau de prix et de qualité du parc de logements. Mais il faut pour cela repérer l'unité de la grille qui contient les communes les plus chères : (0,0) pour la classe 1, (3,0) pour la classe 2, (3,4) pour la classe 3, etc. Ce noeud « central » est environné de noeuds où se retrouvent des communes parfois très éloignées dans la hiérarchie des prix. Cela est particulièrement net pour la classe 1, où l'on retrouve Saint-Denis et Saint-Ouen dans le noeud (0,4), et pour la classe 2, où l'on retrouve Vitry-sur-Seine et Argenteuil dans l'unité (5,1). Par contre, les noeuds sont en général assez homogènes aussi bien au niveau des prix que de la qualité du parc.

Mais l'interprétation des classes s'éclaire si l'on fait intervenir les deux autres « dimensions » que sont d'une part la densité et le poids du pavillonaire, et d'autre part l'ancienneté du parc et le rythme de construction, mesuré à la fois par le rythme annuel moyen de construction autour de la date de référence (ici la moyenne 1981-1983), et par la part des logements construits depuis le dernier recensement (ici après 1975).

Le classement obtenu est donc plus complexe que celui du 4.3.3, mais néanmoins cohérent. Les urbanisations récentes sont concentrées autour de la classe 8; le pavillonaire à droite de la carte avec des pics au N-E et au S-E de la carte. On le vérifie sur les profils ci-dessous, établis pour une variable représentative de chaque « dimension » : le prix du logement collectif ancien, la part des logements très récents, la part des logements sans confort et la part de logement individuels.

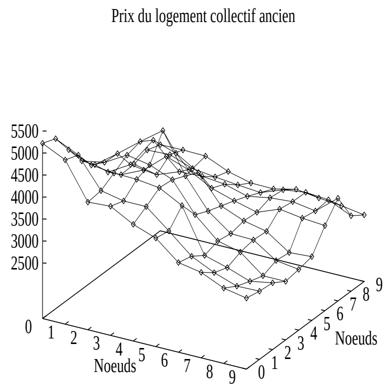


FIG. 4.21 - *La bosse correspond au noeud (3.4) où sont classées les communes de Meudon, Chaville...*

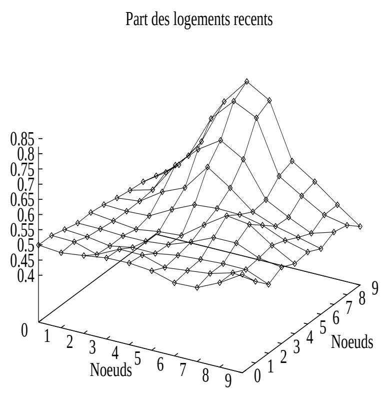


FIG. 4.22 - *Le logement récent se concentre dans la classe 8 et quelques noeuds autour.*

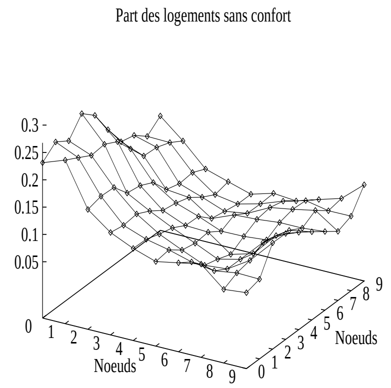


FIG. 4.23 - *Le pic correspond au noeud (0,3) où l'on retrouve Aubervilliers, Saint-Denis...*

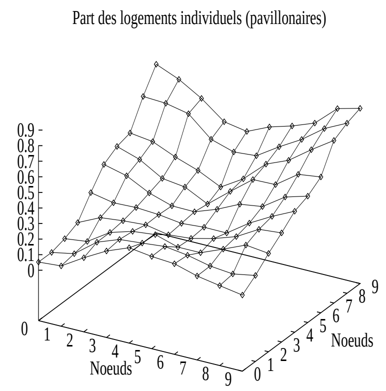


FIG. 4.24 - *Le logement collectif est à gauche de la carte et le logement individuel à droite.*

4.3.5 Commentaires

Après avoir illustré par des exemples de dimension 2 comment l'algorithme reconstitue la topologie des entrées, nous avons montré dans ce chapitre comment on peut utiliser l'algorithme de Kohonen dans le cas de variables quantitatives. La méthode fournit directement une quantification vectorielle (classification) et une projection comparable à celles fournies par une ACP.

Un avantage de l'algorithme de Kohonen est sa rapidité due à la simplicité des calculs. L'algorithme de Kohonen ne fait que des multiplications et des additions (après avoir recherché l'unité gagnante). L'apprentissage du réseau est en général très rapide.

Il faut cependant rappeler que l'objectif n'est pas de substituer l'algorithme de Kohonen aux méthodes classiques d'analyse des données telles que l'ACP et la classification hiérarchique. Il s'agit plutôt d'un complément que nous suggérons d'utiliser pour améliorer, dans certains cas, la visualisation des données et des classes. D'autre part les méthodes classiques sont parfois inadéquates comme dans le cas de bases de données incomplètes, ce qui est le propos du chapitre suivant.

Chapitre 5

Le traitement des données manquantes par KACP

5.1 Les motivations

L'idée d'utiliser l'algorithme de Kohonen dans l'étude DREIF (Cf 4.3.3) nous est venue après avoir constaté que le nombre de valeurs manquantes dans la base de données ne permettait pas de classer une proportion raisonnable de communes. En effet, dans le cas d'une ACP ou d'une CH (Classification Hiérarchique) réalisées avec le logiciel SAS, et en n'utilisant dans le traitement que six variables, seules une trentaine de communes sur les 256 communes que contenait la base restaient dans la table de données. Ce nombre diminue encore si on augmente davantage le nombre de variables puisque à chaque fois que le programme rencontre une valeur manquante, il supprime la ligne correspondante.

Les méthodes classiques sont des méthodes qu'on peut qualifier de méthodes exactes et rigoureuses qui peuvent être inadaptées dans le cas de données incomplètes. Si on regarde le cas de l'ACP, il faut diagonaliser une matrice de variance covariance et de projeter ensuite les individus sur les axes principaux pour construire les composantes principales. Bien qu'il soit possible (notamment par le logiciel SAS) de calculer la matrice de variance covariance en présence de valeurs manquantes, le calcul des composantes principales nécessite, par contre, de faire une estimation des cordonnées manquantes. Cela vaut également pour la classification hiérarchique étant donné qu'il faut calculer un représentant à chaque fois qu'une classe est construite.

La solution « rapide » qui consiste à substituer à la composante manquante la moyenne empirique de la variable correspondante peut être dans la majorité des cas très approximative, et particulièrement dans le cas de grandes bases de données où les variables peuvent avoir de grandes dispersions. L'adaptation faite sur l'algorithme de Kohonen contourne ce problème, car on ne fait aucune estimation préalable sur les données manquantes; l'algorithme utilise uniquement les données qui sont présentes. En fait, cette adaptation pourrait également convenir à un autre algorithme de quantification vectorielle tel que k-means[1] (MacQueen, 1967) mais le choix de l'algorithme de Kohonen est justifié par le fait

1. Il faudrait également voir si l'on obtient les mêmes résultats

que la carte de Kohonen peut aisément remplacer une ACP, en raison de la conservation de proximité, ce qui permet d'interpréter et de mieux distinguer les classes.

Il y a plusieurs raisons qui donnent à cet algorithme cette remarquable robustesse face aux données incomplètes. Ce que l'on peut dire à ce sujet, c'est d'une part le fait qu'il y ait peu de calcul dans une itération de l'algorithme, contribue à ne pas accumuler les erreurs durant les itérations. D'un autre point de vue, il s'agit là d'une méthode neuronale et donc basée sur la notion d'apprentissage. S'il existe une fonction définissant le taux d'apprentissage du réseau, il est naturel de penser que cette fonction dépend de la quantité d'information qui est apportée au réseau. Nous allons tenter de voir comment varie une carte de Kohonen et de quantifier les variations de la distorsion en fonction de la proportion de valeurs manquantes. Nous donnerons à la fin de ce chapitre une méthode d'estimation des données manquantes.

5.2 Principe de la méthode et notations

En analyse des données, une valeur manquante est une observation d'une variable sur un individu qui n'est pas dans la base de données. On suppose que cette observation existe, mais elle n'est pas recensée pour des raisons diverses : erreur de saisie, donnée non disponible, mesure non effectuée, etc...

Si x est un individu de la base Ξ, on appelle M_x le sous-ensemble de $\{1, 2, .., d\}$ qui indique les composantes réellement observées pour l'individu x.

L'adaptation de l'algorithme de Kohonen pour ce traitement est assez intuitive. Dans les deux étapes de calcul de l'apprentissage d'une carte de Kohonen (recherche du minimum et homotétie), on ne travaille qu'avec les composantes présentes. On calcule la distance entre deux vecteurs uniquement sur les composantes observées et de même la mise à jour des des composantes d'un vecteur poids (ou vecteur code) n'opère que sur les composantes qui sont présentes. Les étapes III et IV de l'algorithme défini dans la section 4.1.1 deviennent :

III On recherche l'unité i gagnante en calculant la distance entre une entrée $x \in \Xi \subset \mathbb{R}^d$ et une unité j uniquement par rapport aux cordonnées présentes pour l'entrée x :

$$\|x(t) - y_t^j\|^2 = \sum_{k \in M_x} (x[k] - y_t^j[k])^2,$$

et alors

$$i = \arg\min_j \|x(t) - y_t^j\|^2.$$

IV On modifie ensuite les valeurs des composantes du vecteur code, qui ont le même indice que les composantes présentes de x, pour le noeud gagnant et pour ses voisins par

$$\forall k \in M_x, \quad \forall j \in V_{r(t)}(i), \quad y_{t+1}^j[k] = y_t^j[k] + \epsilon(t)\left(x(t)[k] - y_t^j[k]\right).$$

et pour les composantes non observées on ne fait rien.

On peut se demander alors, comment se fait l'apprentissage d'une composante d'indice k du vecteur code d'une unité i dans laquelle est classé un individu x auquel il manque la k-ème composante?. En supposant que cette k-ème composante ne manque pas à tous les individus de la base Ξ, sinon la k-ème variable n'existe pas, il y aura probablement un autre individu x', dont la k-ème composante est recensée, et qui sera classé dans l'unité i ou dans une unité voisine. De plus au début de l'apprentissage on prend un rayon de voisinage assez grand qui permet de mettre à jour beaucoup de vecteurs codes à chaque itération. En fait la mise à jour de toutes les composantes de tous les vecteurs codes s'effectue sauf dans le cas (à exclure) où toute une variable est absente.

Une des caractéristiques remarquables de cette méthode c'est le fait de ne pas faire d'estimation a priori ou d'interpolation pour compléter les données manquantes. De ce point de vue, on peut dire que cette méthode est très efficace et économique en temps de calcul. En effet, si on prend le cas des données immobilières en Ile de France (voir 4.3.3) où nous avions à établir une classification des communes, pour utiliser les méthodes classiques d'analyse de données, il est nécessaire d'estimer (ou d'interpoler) environ trente pour cent de valeurs sur une base de données constituée des observations, sur 16 années, de quarante variables et 256 communes. Il est clair que cela aurait été beaucoup plus long à réaliser que l'algorithme de Kohonen adapté aux valeurs manquantes.

La méthode que nous avons défini dans le paragraphe précédent (5.2) permet donc d'une part de faire l'apprentissage d'un réseau de Kohonen même avec des proportions de valeurs manquantes « assez grandes ». De plus la qualité des classes obtenues sur les cartes de Kohonen est très satisfaisante, permettant d'avoir une classification et une carte auto-organisée de type ACP, ce qui n'était pas réalisable avec les méthodes classiques. De plus, il est possible d'estimer par cette méthode, et avec une bonne précision, les données manquantes. Ces deux points sont montrés sur des exemples empiriques dans les deux sections suivantes.

5.3 Validation empirique

5.3.1 Mesure d'erreur

Si on suppose que X est une variable aléatoire d-dimensionnelle dont les réalisations sont les éléments de Ξ, on note X^p le vecteur aléatoire issu de X auquel on enlève une proportion p de composantes à chacune de ses réalisations. On fait l'apprentissage du réseau avec le nouvel ensemble Ξ^p des individus auquels on a enlevé une proportion p de composantes. Etant donné le réseau de Kohonen R^p qu'on a entraîné avec la base incomplète Ξ^p, on note q^p le quantificateur vectoriel associé, c'est-à-dire l'application qui à X associe le vecteur code de l'unité gagnante du réseau.

On définit par rapport à ce réseau de Kohonen R^p, entraîné avec une base incomplète, une distorsion qu'on appellera l'«erreur a posteriori» par

$$E^p = \mathbf{E}[\|X - q^p(X^p)\|],$$

qui est l'espérance de la distance entre le vecteur X et le vecteur code associé à X^p par q^p. On calcule les vecteurs codes au cours de l'apprentissage avec les données incomplètes, et on classe les vecteurs incomplets en leur associant leur vecteur code. On regarde ensuite l'erreur moyenne entre le vecteur X avec des données complètes et le vecteur code. Tout se passe comme si après avoir fait l'apprentissage de la carte de Kohonen et après avoir classé les individus de Ξ^p, on retrouvait la possibilité de compléter les données pour permettre alors de calculer l'erreur qu'on a commis en quantifiant Ξ avec $q^p(\Xi^p)$.

Bien entendu cette distorsion est distincte de la distorsion classique du quantificateur q^p qui est :

$$D^p = \mathbf{E}[\|X - q^p(X)\|].$$

Ici on fait l'apprentissage du réseau avec une base de données incomplètes, mais on classe les individus complets et on calcule l'erreur moyenne entre le vecteur d'entrée complet et le vecteur code. Evidemment aucune de ces deux distorsions n'est calculable dans un cas réel possédant des données manquantes. La seule vraie distorsion calculable dans ce cas est alors

$$D' = \mathbf{E}[\|X^p - q^p(X^p)\|].$$

Mais nous n'utiliserons pas cette erreur dans l'étude que nous menons ici; car lorsque l'on augmente la proportion de valeurs manquantes, cette erreur a tendance à baisser. En effet si l'on prend le cas extrême où il ne reste q'une seule composante dans chaque individu, l'algorithme de Kohonen en faisant l'apprentissage du réseau uniquement par rapport à cette composante permet en même temps de faire décroître l'erreur D'. Mais en réalité, le classement des individus, uniquement par rapport à cette composante qui leur reste, est très approché voir désordonné.

Nous allons tenter de vérifier sur des exemples des deux sections suivantes la robustesse de l'algorithme de Kohonen par rapport à l'existence de données incomplètes. Pour cela on considère des bases de données complètes auxquelles nous retirons une proportion variable p d'observations, pour ensuite effectuer l'apprentissage d'un réseau de Kohonen. Nous essayons d'apprécier et de quantifier les déformations du réseau et les écarts par rapport à un état de référence qui est celui où toutes les observations sont présentes.

Nous procédons à deux validations; la première validation concerne la proportion de valeurs manquantes par rapport au nombre de variables, c'est-à-dire qu'on enlève à chaque individu la même proportion p de valeurs observées. Pour la deuxième validation, on fixe le nombre de composantes à supprimer pour une proportion r d'individus. On regarde dans les deux cas, comment varient les distorsions définies ci-dessus.

5.3.2 Validation empirique par rapport aux variables

Les deux premiers exemples ci-dessous sont construits à partir d'un échantillon gaussien de points de $\mathbb{R}^{10}$. Dans les deux cas, la matrice de corrélation des dix variables est distincte de la matrice identité de manière à se trouver dans un cas semblable aux cas réels, puisque comme on l'a vu dans le chapitre précédent, l'organisation respecte les corrélations entre variables. En effet lorsque l'on regarde les profils des variables dans les exemples du chpaitre précédent, on constate que l'algorithme de Kohonen organise la carte en tenant compte des corrélations de variables. Le troisième exemple est un cas simulé à partir de la base des données immobilières et démographiques d'Ile de France.

Echantillon en trois classes

Dans cet exemple, on a un échantillon de points de $\mathbb{R}^{10}$ constitué de trois classes linéairement séparables et désignées par «1», «2» et «3».

Chaque classe contient un effectif de 200 individus et on enlève à tous les individus, successivement une composante sur les dix, puis deux et ainsi de suite jusqu'à neuf. Les composantes supprimées sont choisies d'une manière aléatoire. On a donc dix bases de données $\Xi^0 = \Xi$, Ξ^1,..., Ξ^9 avec lesquelles on fait l'apprentissage d'un réseau de Kohonen 8×8. Ce qui donne dix cartes selon la proportion p de valeurs manquantes qui va de 0 à 90 pour-cent.

De plus, pour les dix cas ($p = 0\%$ à $p = 90\%$), on a réalisé 300 versions de cartes de Kohonen et pour chacune de ces versions, on a calculé l'erreur a posteriori E^p et la distorsion D^p. On a donc deux fonctions $E^p(v)$ et $D^p(v)$ de l'indice de version v où $1 \leq v \leq 300$. Les figures 5.1 et 5.1 représentent les variations de E^p pour les 300 versions et différentes valeurs de p.

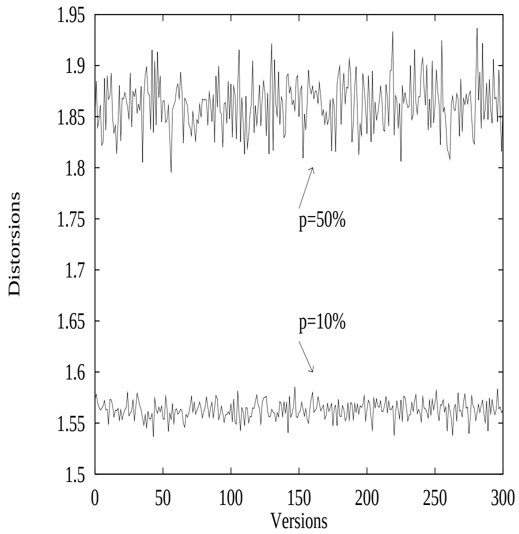


FIG. 5.1 - *Les variations de l'erreur sont comparables pour $p = 0\%$ et $p = 50\%$*

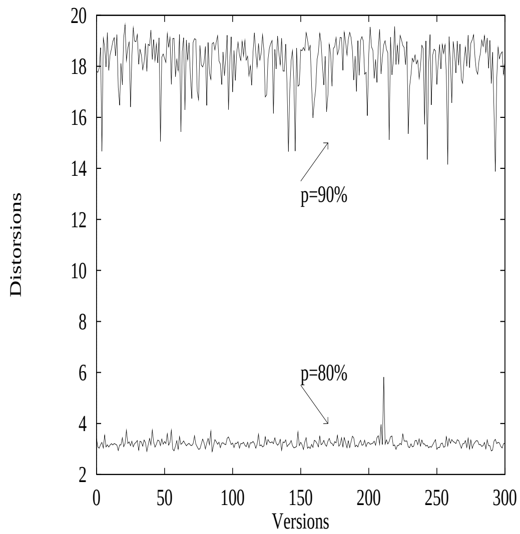


FIG. 5.2 - *Pour $p = 80\%$ et $p = 90\%$ la moyenne et les variations sont plus importantes*

On remarque donc, pour toutes les valeurs de p, que l'erreur E^p varie autour d'une valeur moyenne constante. D'autre part, plus la valeur de p augmente, plus la variance de $E^p(v)$ augmente. On peut noter également que l'écart type de E^p pour $p = 90\%$ est beaucoup plus grand que pour $p = 80\%$ ou pour $p = 50\%$, et $p = 10\%$...

Dans la figure 5.3 sont représentées, pour p variant de 0 à 90, les moyennes $\bar{E^p} = \mathbf{E}(E^p(v))$ et $\bar{D^p} = \mathbf{E}(D^p)$ par rapport à v. On s'aperçoit que $\bar{D^p}$ est toujours inférieure à $\bar{E^p}$. Les deux courbes restent quasi plates jusqu'à $p = 50\%$.

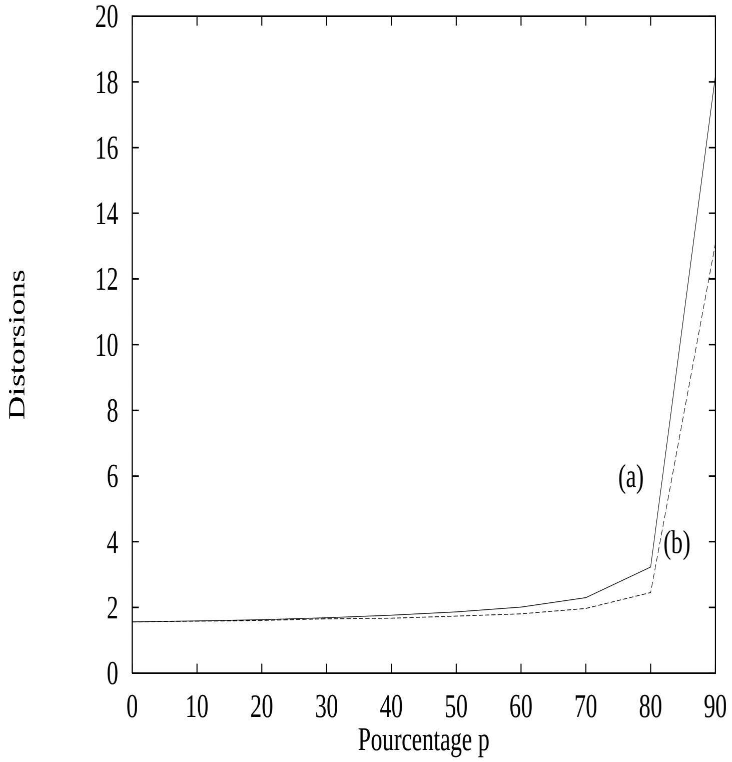


FIG. 5.3 - *(a) Erreur a posteriori moyenne $\bar{E^p}$, (b) distorsion moyenne $\bar{D^p}$.*

Les cartes de Kohonen des tableaux 5.1 à 5.3 confirment cela puisque les trois classes restent bien séparées jusqu'à $p = 70\%$. Les premiers mauvais classements apparaissent quand $p = 80\%$ et la carte est totalement désorganisée pour $p = 90\%$.

3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3		2 2 2 2 2 2	2 2 2 2 2 2	2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3		2 2 2 2 2 2 2 2 2	2 2 2 2	2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3		2	2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2	2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3			2 2 2 2 2 2 2 2 2 2 2 2 2	2 2		1
	2 2 2	2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2	2 2 2 2 2 2 2	2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2	2 2 2 2 2	2		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

TAB. 5.1 - $p = 0\%$, *les trois classes sont bien séparées*

2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2	3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
2 2 2 2 2	2 2 2	2 2 2 2		3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2	2 2 2 2 2 2 2 2		3	3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
	2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2			3	3 3 3 3 3
1 1 1 1 1 1 1 1	1 1		2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2	2 2 2 2 2	2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1		2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1		2 2 2 2 2 2 2	2 2	2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	2	2 2 2 2 2 2 2 2 2 2	2 2 2 2	2 2 2 2 2 2

TAB. 5.2 - $p = 70\%$, *les trois classes sont encore bien séparées*

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3	1 1 2 2 2 3 3 3	1 1 1 1 1 1 2 2 2 2 2 3 3 3	1 1 1 2 2 2	1 1 1 1 2 2 2 2 2 2 2	1 1 1 1 1 2 3 3 3	1 2 2 3 3 3 3	1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 3
2 2 2 2	1 1 1 1 1 1 1 2 2 2 3 3 3	1 1 2 2 2 2 2 3 3 3	1 1 1 1 1 2 2 2 2	1 2 2 2	1 1 1 2 2	1 2 2 2 2 3 3 3 3	1 1 1 1 1 2 2 2 2 3 3
1 1 1 1 1 1 1 1 1 2 2 2 3	1 1 1 2 2 2 2 2 2 3 3	3	2 2	1 2 2 2 2 2	1 2 2 2 2 2	1 1 1 1 1 1 1 2 2 3	1 1 2 2 2 3 3 3
1 2 2 2 2 3	2 2 2 2 2	1 2 2 2 2 2 2	2 2 2 2	2 2 2 2 2 2	2 2 2	1 2 2 2 2 2 2 2 3	1 1 1 1 1 2
1 1 1 2 2 2 2 2 3	2 2 2 2 2 2 3 3	2 2 2 2 2	2 2 2 2	2 2	2 2	1 1 2 2 2 2 2 2 2	1 1 2 3 3
1 1 1 1 1 1 2 2 2 2 2 3 3 3	1 1 1 1 1 1 1 2 2 2 3 3 3 3 3 3	1 1 2 2 2 2 3 3 3		1 2	1 1 2 2 3 3 3 3 3 3 3 3 3 3 3	1 1 2 3 3	1 1 1 2 2 2 2 2 2 2 2 3
2 2 3 3 3 3 3 3	1 1 1 3 3 3 3 3	2 2 3 3 3	1 1 2 3	1 2 2	2 3 3 3 3 3 3	1 1 1 1 1 1 3 3 3 3 3 3 3 3	1 1 2 3 3 3 3 3
1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	1 1 2 3 3 3 3	2 3 3 3 3 3 3 3 3 3 3 3 3 3	1 2 2 2 3 3 3	1 1 2 2 3 3 3 3 3 3	1 1 1 1 1 1 2 2 3 3 3 3 3 3	1 1 1 1 2 2 2 3 3 3 3 3 3 3	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

TAB. 5.3 - $p = 90\%$, *les trois classes se mélangent totalement*

Echantillon en quatre classes

Dans cet exemple, on a un échantillon de points de $\mathbb{R}^{10}$ constitué de quatre classes composées chacune de 100 individus. Les quatre classes symbolisées par «T», «X», «O» et «Z» se recouvrent partiellement (il n'y a pas de recouvrement simultané de trois classes). En effet si on regarde la carte de Kohonen 5.4, on ne trouve aucune unité contenant simultanément les trois classes. La classe «Z» intersecte la classe «X», celle-ci intersecte aussi la classe «O» et la classe «T» ne touche que la classe «O».

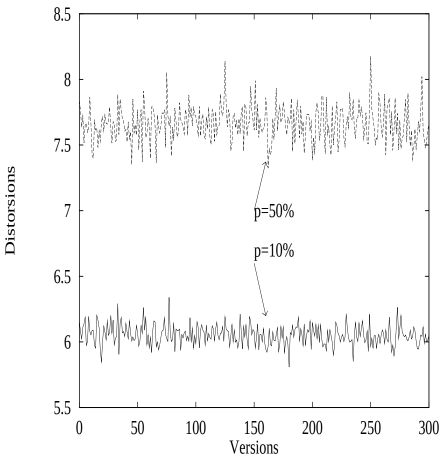


FIG. 5.4 - *E^p pour $p = 0\%$ et $p = 50\%$; les variances sont peu différentes.*

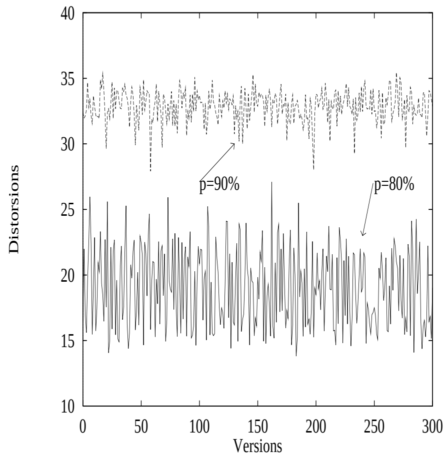


FIG. 5.5 - *La variance de E^p pour $p = 80\%$ est supérieure à celle pour $p = 90\%$.*

Pour les figures 5.4 et 5.5, on peut reprendre les remarques faites pour l'exemple précédent, à l'exception du fait que la variance de E^p baisse brusquement pour $p = 90\%$.

D'autre part, on constate aussi que la courbe de la distorsion moyenne $\bar{E^p}$ croît très faiblement jusqu'à $p = 50\%$. La première carte de Kohonen représentée par le tableau 5.4 montre les classes qui, bien qu'elles se «touchent», sont facilement distinguables. Pour $p = 70\%$, on voit (tableau 5.5) que les classes se recouvrent davantage, puisqu'il y a des unités qui contiennent des individus appartenant aux trois types de classes. Lorsque $p = 90$ (tableau 5.6), elles sont entièrement mélangées.

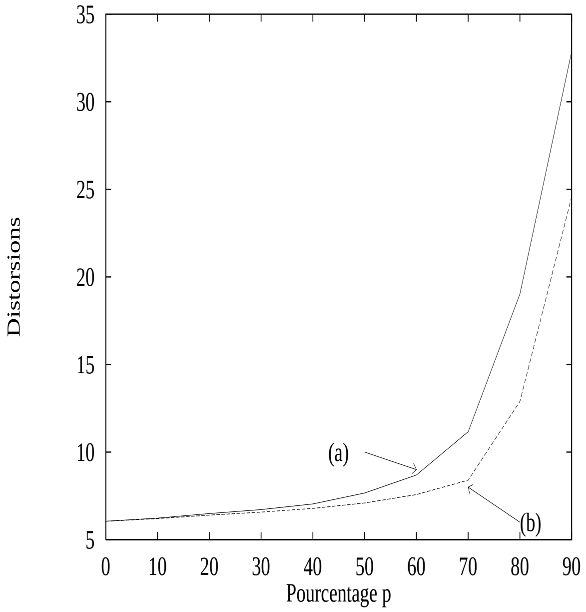


FIG. 5.6 - *(a) Erreur moyenne* $\bar{E}^p$*, (b) Distorsion moyenne* $\bar{D}^p$*; les courbes sont également quasi plates jusqu'à* $p = 50\%$

T T T T T T	T T T T T T T	T T T T T T T T T T T T	T T T T T T T T T	T T T T T T T T T T T T T	T T T T T T T T T T T T	T T T T T T T T T T	T T T T T T T
O O T	T	O O T	O T T T	T T T T T	T T T T T T T	T T	T T T
O O O O O O	O O O O	O O O O O O O O O O O O O T	O O O O O O	O O O O O O O O O	O O O O O O O O O O O O	O O O O	O O
O O O O	O O *X*	O O O O O	O O O O *X*	O O O O O	O O O O O O	O O O O O O *X X*	O O O
X	*X X X*	O *X X X* *X*	O O *X X* *X X X X* *X X*	*X X X X* *X*	*X X X X*	*X X X X*	O *X X X* *X X*
X X X X *X X X X* *X*	*X X X X* *X*	*X X X X* *X X X X* *X X*	*X X X X* *X X X X* *X*	*X X X X* *X X*	*X X X X* *X X* Z	*X X X X* *X X X* Z	*X X X X* *X X X*
X X	Z	Z Z Z Z	Z Z Z Z	Z Z	Z Z Z	*X* Z	Z Z Z
Z Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z Z Z Z Z	Z Z Z Z Z Z Z Z Z

TAB. 5.4 - $p = 0\%$, *les classes sont séparables*

X Z Z Z Z Z Z Z Z Z Z	Z Z	X X X X X X X	O O X X X X X X X X Z	O O O X X	O O O O O	T T	O T T T T T T T T T T T T T
Z Z Z Z Z Z Z Z Z Z	X X Z	O O X X X	O X X X	O O O O O X	O O O O O O T X X	O O T	T T T T T T T T T
X Z Z Z Z Z Z Z Z	X Z	X X X X Z Z	X X	O O X	O O O O O O O T	O	T T T T T T T T T T T T T T T T T
Z Z Z Z Z Z Z Z Z Z Z Z	Z	X X X X X X Z	X X X X	O O O O O O O Z	O O O	O O T T	T T T T T T T T
Z Z Z Z Z Z Z Z Z	Z Z Z Z	O X X X X X X	O T X X X	O X X X	O O O O O O O O O O O	O O	T T T
X Z Z Z Z Z Z Z Z Z Z Z Z Z	X Z Z Z Z	X X X X	O O X X	O O O X	O O O O O O T T	O T	O T T T T T T T T T T T T T T T T
Z Z Z Z	Z	X X X X X X X X X	O O X X X	O X	O O X	T T T T	T T T T T T T
X Z Z Z Z Z Z Z Z Z Z Z	X X X Z Z Z	X X X X X X X X	O O X X X X	O O O O O O O X Z	O O O O O O	O O T T T	T T T T T T T T T

Tab. 5.5 - $p = 70\%$, *les classes se recouvrent davantage*

T T T T T T T T T T X X X X X X X X Z Z Z Z Z Z Z	O T Z	O O O T X X X Z Z Z	O X	T Z Z	O O O T X Z	O T T X Z	O O T T T T T T X X X X X Z Z Z Z Z Z Z Z Z Z
X	O O O T T X X Z Z	T Z Z	O T T	O O O T T X Z	O O Z Z Z Z Z	X X X Z Z	X Z Z Z Z
O O O T T Z Z	O T T Z Z Z	O O T T T X Z Z	X X X X	O X X X Z Z Z	O O X Z Z Z	O X X X X Z	O O O O T T X Z Z Z Z Z
O O O X X X X X Z Z Z	T X X Z	O T X	O O T Z	O X	O O O O O T T X X Z	O O T T X Z	O X Z Z
O O Z Z Z Z	O O O T X Z Z Z	O O X X	X	T	O T X X	O O X X	O O X X X X Z Z
O T T X	O X	O O T T Z	O O O X	O O T	O O X Z	O X	O O T X X X
O O O O T T T T Z	O T T Z Z	O O O T T X X X Z Z	O T T X	O T X X Z	O T X X X Z Z Z	T T X Z Z Z	O T Z
O O O T T T T T T T X X Z Z	O T X X X	O T T T T T T X	O O O T T T X X X X Z Z	O O O T T	T T T X	O T X X Z	O T T T T T T T T X X X Z Z

Tab. 5.6 - $p = 90\%$, *les classes se mélangent totalement*

Un cas réel; les données démographiques en Ile de France

Il s'agit d'un extrait du tableau utilisé dans l'étude sur le marché du travail en Ile de France. On a à l'origine, un tableau qui contient 298 individus et 22 variables et on regarde comme précédemment, les distorsions E^p et D^p en fonction de la proportion p de valeurs manquantes.

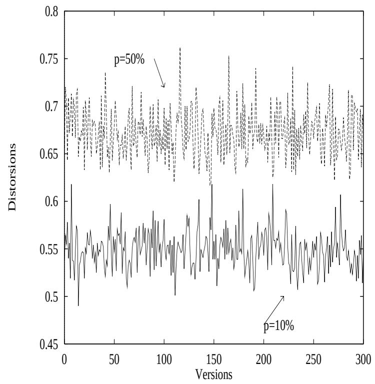


FIG. 5.7 - *L'erreur E^p pour $p = 0\%$ et $p = 50\%$.*

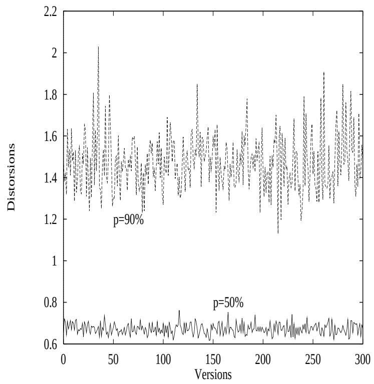


FIG. 5.8 - *L'erreur E^p pour $p = 50\%$ et $p = 90\%$.*

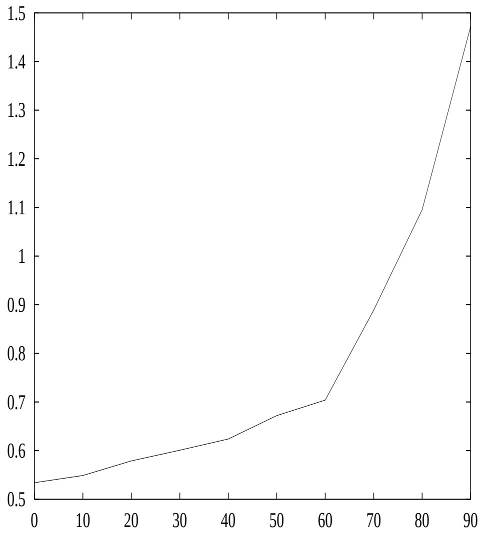


FIG. 5.9 - *L'erreur moyenne $\bar{E}^p$ en fonction de p; ici elle croit plus rapidement que dans les deux exemples précédents.*

L'échantillon ici est plus complexe et pour mieux voir les mouvements de classes on a fait un classement préalable[2] en huit segments. Le nom de chaque commune est précédé du numéro du segment auquel elle appartient. Ainsi la classe six est constituée uniquement

2. C'est le résultat d'une Classification Hiérarchique

des arrondissements de Paris (pa1 ... pa20). On prend cette classe comme repère pour évaluer les désorganisations car il se trouve que c'est la classe qui se distingue le plus des autres classes. Mais rien ne nous empêche de voir aussi les autres classes !.

La figure 5.9 montre que la moyenne de l'erreur $\bar{E}^p$ croit en fonction de p d'une façon plus rapide que dans les exemples précédents. C'est dû, certainement, au fait que la base de données est plus complexe. Il est très probable qu'on n'ait pas comme précédemment, des classes qui soient bien séparées dès le départ où qui ne se recouvrent que partiellement.

On remarque que jusqu'à $p = 50\%$ (tableau 5.8) la classe six qui est composée exclusivement des arrondissements de Paris, reste bien séparée des autres classes, excepté pour le 19^{eme} arrondissement (pa19) qui se retrouve affecté à l'unité (4,8). De même pour les classes huit et deux qui restent bien homogènes jusqu'à $p = 50\%$.

Cependant pour les classes qui au départ se recouvrent déjà, telles que les classes un et cinq, ou les classes quatre et trois, les déclassements apparaissent un peu plus tôt, c'est-à-dire qu'elles se recouvrent davantage. La carte est pratiquement désorganisée à partir de $p = 80\%$ (tableau 5.10).

Comme conclusion de ce premier volet de cette étude empirique, on constate donc que si les classes sont bien séparées, alors elles peuvent le rester jusqu'à une proportion p avoisinant les soixante-dix pour-cent. Par contre au sein des classes, il y a évidemment des mauvais classements. Cela se voit lorsque l'on a des classes qui se recouvrent.

4GARV 4HOUD 3MIMO 4ORME 4STPI 3VILP 3VROI	1MONG 4PAVB	7BRYM 7EPIO 7IGNY	7BURE 7GIFS 5MARN 7ORSA 7VERB	5SCAU 5VAUC 5VAVR 5VESI	5NEUS 5SCLO	6pa15	6pa01 6pa02 6pa04 6pa05 6pa06 6pa07 6pa08 6pa09 6pa16
4ARNO 4AUVE 4BEAU 4MORA	3CHEL 4CONF 4GROS 4HOUI 4ISLE 4LIVR 3MORS 4NEUP 4SAVO 3STFA	7EAUB 7PALA	5CELL 5CHAT 5CHES 5MARL	5BGLR 5GARC 5MAIL 5SEVR 5VIRO	5ENGH 5SGER 5SMAN 5VERS		6pa03 6pa10 6pa11 6pa12 6pa13 6pa14 6pa17 6pa18 6pa19 6pa20
7AUFF 7CROS 7ESSA 4HERT 7LESI 7MENN 4PERY 7SLEU	7CORM 4MSON 4SUCY 7YERE	4BRUN 4DEUI 4DRAV 4PERX 4VAIR 1VMOM	5ANTO 5FONR 5MEUD 5PECQ 5RUEI	5CHAV 5MMOR 5NOGM 5RAIN 5SMAU 5VANV	1BOIC 5BOUL 5CBVE 1CHTN 1GARE 1ISSY 4JUVI 1MRGE 5VINC	1FONT	1LEVA
7ANDR 3DOMO 7MAUR 7VIPE	7VERN	1AVON 3RAMB 1SOIS	5CHON 1HAYR 5VELI	1CACH 1CLAM 1FONB 1JOIN 1MALF	1ACRU 1BOUR 1KREM 5SMOR 1SURE 1VJUI	1ASNI 1COLO 1LILA 1MALA	1PUTO
3OSNY 3PORC	3ELAN 3ERAG 3PLAI 3TAVE	3BOIA 3CHIL 3CLAY 5FONF 3FRAN 3SMIC	3ERMO 3LIME 3LONG 3RISO	1CHEV 1CRET 1FRES 1MASS 1NEUM 1PTOI 1SCYR	1BGNX 5CHMA 1PLES	2CHOI 2GENT	2ALFO 2CLIC 2IVRY 2MTRE 2PANT 2PREG 2SOUE
3COMB 3DAMG 3OZOI 3PONT 3ROIB	3BRIC 3RUNG	3PLET 3SANN	3BRET 1GRAT 3RONY 3THIA 3VIRY	2GONE 2NOIS	2ACHE 2BOND 2EPIS 2EPNS 1POIS 2SARC 2SVRA 1ULIS 2VBEL	2ISTD 2NANT 2VITR	2AUBV 2BAGT 2ROMA 2SDEN 2VSTG
8JOUY 8MONB 8SACL 8VOIS	8SAVI	3CHEN 3SART	3ATHI 3BRIE 3CARS 3GAGN 3LAGM 3SGEN 3VILM	2ARPA 3DOUR 3MANV 3VIGS	2DAML 2ETAM 2MELU 2OUEN	2CORB	*2BOBI* *2BONE* *2GENN* *2LACO* *2ORLY* *2STAI* *2VALE* *2VGAR* 2VTAN
8CERG 8COUR 8LOGN 8MOIS 8ROIF 8VAUR	3CHSM 3EVRY 3GUYA 3MEES 3TORC	3AUBG 3BOIL 3COIG 3MCOR 3MMAG 3NOIG 3NOIL 3VPIN	3TREF	2ARGE 2AULN 2BEZO 2BLAN 2CHPY 3DRAN 2GOUS 2MFER 2PIER	3COUL 3LIMA 3MAGN 2NEMO 2PERS 2PROV	2MEAU 2MONT	2CARP 2CHAN 2CLIB 2GARG 2GRIG 2MANJ 2MURO 2TRAP 2VERI

TAB. 5.7 - *p=0 %, la classe 6 (Paris) est bien séparée*

6pa01 6pa02 6pa03 6pa04 6pa05 6pa06 6pa07 6pa08 6pa09 6pa14 6pa16 6pa17 6pa18	6pa12 6pa15 6pa20	5MARN 5SCAU 5SCLO 5SMAN	7BURE 7GIFS 7VERB 5VESI	7AUFF 7ORSA 5VERS 5VIRO	7CORM 5FONF 7IGNY 7LESI 8SACL 7SLEU	4BEAU 3DAMG	4ARNO 4AUVE 7EPIO 7ESSA 4GARV 4HERT 4ORME 4PERY 4STPI
6pa10 6pa11	6pa13	5BGLR 5ENGH 5GARC 5NOGM	5MAIL 5SMAU	5MMOR 5RAIN 5VAUC	4BRUN 5CHMA 4DEUI 7EAUB 4LIVR 7MENN 4MSON 4SAVO 4VAIR	7CROS 4DRAV 4MORA	3MORS 3PONT 7YERE
5VAVR	5CBVE 5NEUS 5SGER	5BOUL 1LEVA 5VINC	5CHAV 1CHTN 5MEUD	4ISLE 4PAVB 4PERX	1AVON 3SANN 1SOIS	3BOIA 4CONF 3FRAN 3RAMB 3SMIC 4SUCY	3AUBG 3BRIC 3BRIE 3CHEL 4GROS 3LIMA 4NEUP 3OZOI 3PLET 3PORC 3STFA 3VILP
5CELL 5CHAT 5CHES 5FONR 5MARL 5PECQ 8ROIF 3RUNG		1BOIC 1GARE 1ISSY 5SMOR 5VANV	1ASNI 1CLAM 1COLO 4JUVI 1MALA 1MRGE	1CACH 1HAYR 1JOIN 1LILA 1MALF 1MONG 1VMOM	2BLAN 3GAGN 1GRAT 4HOUI 7PALA 3RONY 3VIGS 3VIRY	3BRET 3LAGM 3SART 3SGEN 3VILM	2ARPA 3CARS 4HOUD 3MANV 3MIMO
5ANTO 5VELI 7VIPE	5CHON 5SEVR	1BOUR 5RUEI	1FONB 1KREM 3LIME 1PTOI 1SURE	1VJUI	3ATHI 2CHOI 2NOIS	2ARGE 2CHPY 3VROI	3COUL 3DOUR 2ETAM 1FONT 3MAGN 6pa19
7BRYM 3COMB 3ELAN 8MONB 3PLAI	7ANDR 3DOMO 3ERAG 3TAVE 7VERN	3CHIL 3ERMO 7MAUR 3RISO	1ACRU 1CHEV 3CLAY 1FRES 1POIS 1SCYR 1ULIS	1BGNX 1PUTO	2ALFO 2AUBV 2BEZO 2GENT 2IVRY 2MTRE 2PANT 2PREG 2SOUE	2BOND 3DRAN 2PIER 2VITR 2VSTG	2AULN 2GOUS 2MEAU 2PROV
8VOIS	3OSNY 3ROIB 3TREF	3COIG 3LONG 3MCOR 3THIA 3VPIN	3CHEN 1CRET 1MASS	2CORB 2EPNS 2SARC 2VGAR	2BAGT 2CLIC 2NANT	2ISTD	2ACHE 2NEMO 2PERS 2VTAN
8CERG 8COUR 8JOUY 8LOGN 8MOIS 8VAUR	3GUYA 3MEES 8SAVI 3TORC	3BOIL 3CHSM 3EVRY 3MMAG 3NOIG 3NOIL	2CARP 2EPIS 1NEUM	2DAML 1PLES 2SVRA 2VBEL 2VERI	2BONE 2CHAN 2CLIB 2GARG 2GRIG 2MELU 2MURO 2OUEN 2TRAP 2VALE	2BOBI 2GENN 2LACO 2MANJ 2MONT 2ORLY 2STAI	2GONE 2MFER 2ROMA 2SDEN

TAB. 5.8 - *p=50%, la classe 6 est toujours séparée*

7AUFF 7CROS 7LESI 3MIMO 4MORA 4ORME 3SMIC 4STPI	3BRIE 4GARV 4GROS 3STFA	4ARNO 3AUBG 3CHEL 2CHOI 1HAYR 2MFER 3MORS 3RONY 3SGEN 3TREF 3VILM	1CLAM 3DRAN 4HOUD 2PERS	2ALFO 2BAGT 2GOUS 3LIMA 2MONT 2VITR	2AULN 2BEZO 3CARS 2GARG 2NOIS 2PIER 2SOUE	2CHAN 2CLIC 2MURO	2BOBI 2BONE 2CORB 2EPIS 2ORLY 2PANT 2PREG 8ROIF 3RUNG 2VGAR
4AUVE 3CHIL 7CORM 7EPIO 4PERY 7SLEU	4NEUP 4SAVO	3ATHI 2IVRY	3MANV	2PROV 3VROI	2CHPY 2MTRE 2SDEN	2ARGE 2BOND 2ISTD 2MELU	2ACHE 2MANJ 2SVRA 2VALE 2VBEL
5ENGH 4HERT 4LIVR 4MSON 4PERX 3PLET 4SUCY	4BEAU 4BRUN 4CONF 4HOUI 4PAVB	1COLO 4DEUI 4DRAV 3GAGN 1SOIS 1VMOM	3LAGM 3MAGN	2ARPA 3CHEN 3COUL 2NEMO 3VIGS	2LACO 2NANT 2ROMA 2VSTG	2DAML 2SARC	2AUBV 1BGNX 2EPNS 2GRIG 2STAI 2VERI 3VILP 3VIRY
7BURE 7EAUB 5MARL 7YERE	1AVON	4ISLE 1MONG 3RAMB 4VAIR	1CACH		1BOUR 2GENN 1MALA 1PUTO	1ULIS	2BLAN 2CARP 2CLIB 2MEAU 1PTOI 3VPIN 2VTAN
7GIFS 7IGNY 7ORSA 5SCAU 5VAUC 5VAVR 5VIRO	5BGLR 5CHAT 5GARC 5MMOR 5SMAU	5CHON 1GARE 1JOIN 5MAIL 7MENN 7PALA	5CHMA 1FONB 5MEUD	1KREM 1LILA 3LIME 1POIS 1VJUI	2ETAM	3BRET 3COIG 1GRAT 3PONT	3BOIL 1CRET 3DAMG 3EVRY 2GONE 3MEES 3SART
1MRGE 5VESI	5MARN 5SEVR 7VERB	1BOIC 5CHAV 5NOGM 5RAIN 5SCLO 5SGER 5VERS 5VINC	2GENT 4JUVI 1SURE	1MALF 1MASS 3PORC	1CHEV 3CLAY 3DOUR 1FONT 5PECQ 1PLES 3RISO 3SANN 1SCYR 3TAVE	3BOIA 1FRES 3LONG 3MCOR 1NEUM	8CERG 8COUR 3MMAG 8MOIS 3NOIL 3OSNY 2OUEN 3PLAI 3TORC
6pa05 6pa13 6pa14 6pa16	5NEUS 6pa15		7BRYM 5CHES 5VELI	5ANTO 5SMOR 7VERN	3ELAN	3BRIC 3ERMO 3FRAN 3OZOI	3DOMO 3NOIG 8SAVI 3THIA 2TRAP
6pa01 6pa02 6pa03 6pa04 6pa06 6pa07 6pa08 6pa09 6pa10 6pa11 6pa17 6pa18 6pa20	6pa12	5BOUL 1CHTN 1LEVA 6pa19	1ASNI 5CBVE 5CELL 5FONR 1ISSY 5SMAN 5VANV	1ACRU 7ANDR 7ESSA	3ERAG 5FONF 7MAUR 5RUEI	3GUYA 7VIPE	3CHSM 3COMB 8JOUY 8LOGN 8MONB 3ROIB 8SACL 8VAUR 8VOIS

TAB. 5.9 - *p=70 %*

3RUNG 5VAUC 6pa02 6pa06 6pa16 6pa17	5NEUS 5SCAU 6pa05 6pa07 6pa08	5SCLO 6pa04 6pa13 6pa14	5CELL 2EPIS 5SMAN 5VINC 6pa03 6pa12	5BOUL 1CHTN 6pa11 6pa20	1CACH 2GENT 1ISSY 2IVRY 4JUVI 1LEVA 1PUTO	1ASNI 2LACO 3VILP	5GARC 1LILA 2MANJ 2MURO 1VJUI
1AVON 5VESI 3VROI 6pa01 6pa15	6pa09	7ANDR 5SMOR	1FONT 5PECQ 7VERB	1MALF 5NOGM 2ORLY	2BEZO 5CHON 1MALA 1MASS 1SURE 6pa19	2CHOI 2ISTD	2BAGT 1COLO 1GARE 2GRIG 2MFER 2NANT 2ROMA 2VGAR
7GIFS 1SOIS 5VIRO 6pa10	5ENGH 4PERX	5BGLR 4DEUI 5MAIL 5RAIN 5SEVR 5VERS	1BOIC 5CHAT	1BOUR 4DRAV 1VMOM	1CHEV 2PREG 2SARC 3THIA 3VIRY	3MAGN 2PANT	1FONB 2MELU 2SOUE 2VSTG
5CHAV 7IGNY 5MEUD 7ORSA 5VAVR	7BRYM	3CHEL 5CHES 5FONR 3GUYA 5MARL 4MORA 1MRGE	1CLAM 3ERAG 1JOIN 5MMOR 2MTRE 5SGER	3BOIA 5CBVE 4GROS 1KREM 3LAGM	3ATHI 2CLIC 4NEUP 1POIS 1PTOI		2ACHE 2AUBV 2STAI 2VITR
4AUVE 4BEAU 7EPIO 4HOUD 7MENN 4PAVB 4VAIR 5VELI		5ANTO 7BURE 7MAUR 5VANV	7CORM 4SAVO	4CONF 2GOUS 1GRAT 3MORS 3RISO 3STFA 4SUCY	1ACRU 3CARS 3CHEN 2ETAM 3MANV 3MIMO	2ALFO 1CRET 3DRAN 2PERS	2BOBI 2CARP 2EPNS 3EVRY 2GARG 1SCYR 2SDEN 2TRAP
7AUFF 7CROS 7EAUB 7ESSA 4MSON 4PERY 7SLEU 5SMAU 8VOIS	2NOIS 7PALA 3PLAI 5RUEI	3DOMO 3OSNY 8SACL 7VERN	3BRET 5MARN 8VAUR 7VIPE	3COIG 1PLES 3SGEN	2GONE 3MCOR 3NOIG 8SAVI 3VIGS	2PIER 3SART 2VERI	3AUBG 2BLAN 2BONE 3LIMA 3LONG 2OUEN
2GENN 7LESI 6pa18	4HERT 4ISLE 4LIVR 7YERE	3CHIL	3PONT 3TAVE	1NEUM	3MMAG 3ROIB 2VBEL 3VILM	2BOND 3CLAY 3DOUR 2SVRA	1BGNX 2CHPY 3LIME 2MEAU 2VTAN
4ARNO 3BRIC 4GARV 2MONT 2NEMO 4ORME 3OZOI 3PORC 4STPI 3TREF	4BRUN 3DAMG 3ERMO 1HAYR 1MONG 3SANN 3VPIN	8COUR 4HOUI 3RONY 3SMIC	2CLIB 3COMB 3FRAN 8ROIF	3CHSM 3ELAN 1FRES 3GAGN 8JOUY 8LOGN 3MEES 8MOIS 8MONB 3PLET	3NOIL 1ULIS	2ARGE 3BOIL 3BRIE 5CHMA	2ARPA 2AULN 8CERG 2CHAN 2CORB 3COUL 2DAML 5FONF 2PROV 3RAMB 3TORC 2VALE

TAB. 5.10 - *p=80 %*

5.3.3 Validation empirique par rapport aux nombre d'individus

Dans cette section, nous fixons le nombre m de composantes à supprimer par individu et on fait croître la proportion d'individus ayant des observations manquantes. Il est clair que ce cas est moins contraignant que le cas précédent, mais ce que l'on cherche à identifier dans ce cas, c'est parmi les deux situations suivantes, celle qui est la plus pénalisante :
1)- on enlève m composantes choisies au hasard.
2)- on enlève toujours les m premières composantes.
L'exemple suivant dans le plan permet d'introduire le problème.

La loi uniforme dans un carré de $\mathbb{R}^2$

L'exemple qui est traité dans cette section a pour but de voir les déformations du réseau en fonction de la proportion d'individus qui ont des valeurs manquantes. On considère un échantillon de 1000 points tirés uniformément dans un carré de longueur 10 et on fait croître la proportion q d'individus ayant une composante qui manque. Mais il y a deux manières possibles d'enlever ces composantes : 1) sur la proportion r d'individus on enlève une composante sur deux d'une façon aléatoire, 2) on enlève à chaque fois la deuxième composante. Afin de différencier les deux cas, on note par l'indice 1 le premier cas et par l'indice 2 le deuxième cas.

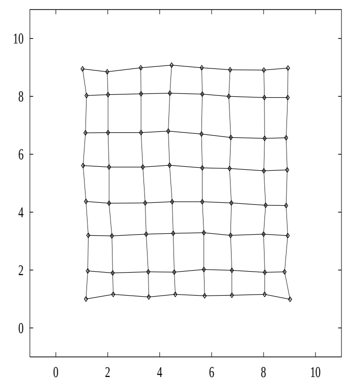


FIG. 5.10 - $r_1 = 0\%$

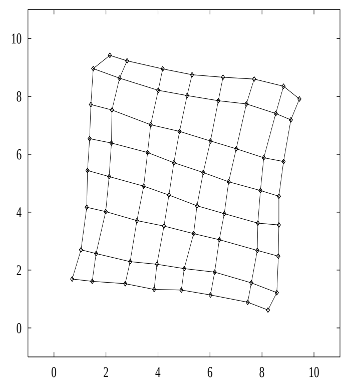


FIG. 5.11 - $r_1 = 40\%$

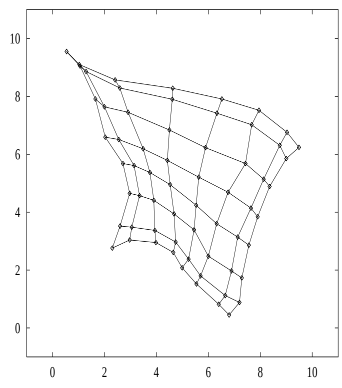


FIG. 5.12 - $r_1 = 90\%$

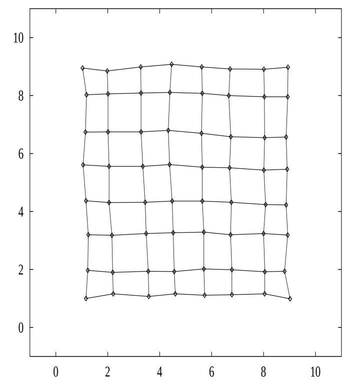


FIG. 5.13 - $r_2 = 0\%$

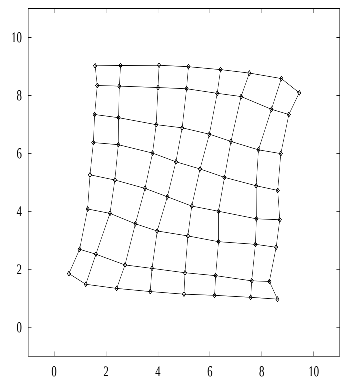


FIG. 5.14 - $r_2 = 40\%$

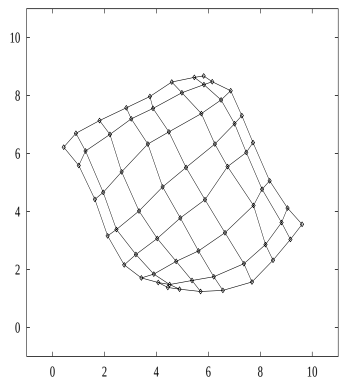


FIG. 5.15 - $r_2 = 90\%$

Les trois premières figures (5.10, 5.11 et 5.12) correspondent au cas où l'on supprime une composante au hasard sur les deux que contient chaque individu. Pour les trois dernières figures, on a enlevé la deuxième composante à chaque individu de la proportion r_2. On remarque alors que dans les deux cas la forme carrée du réseau est assez bien conservée jusqu'à la proportion de 40 pour cent. Au delà les réseaux se déforment mais pas de la même manière. Il y a un effet de rotation dans la deuxième stratégie.

Les exemples de mélanges de classes

Pour les deux exemples traités dans la section précédente, on calcule les erreurs E^p sous les deux hypothèses (1) et (2) décrites ci-dessus, et en fixant à sept (sur 10 composantes au total) le nombre m de composantes à enlever. Précisément pour l'exemple de l'échantillon en trois classes (5.3.2), nous enlevons sept composantes choisies aléatoirement parmi les dix, avec une proportion r_1 allant de 0% à 100% dans le cas (1). Dans le cas (2), on supprime les sept premières composantes avec une proportion r_2 variant de 0% à 90%. Lorsque $r_1 = 100\%$ on est dans la situation de la section précédente avec $p = 70\%$ ce qui, comme on a pu le constater, correspond au début de la désorganisation. On procède aux mêmes traitements pour l'exemple constitué de quatre classes (5.3.2). Les courbes de résultats ci-dessous sont obtenues en faisant la moyenne sur 300 versions de cartes de Kohonen constituées de 8×8 unités.

Cas de trois classes

Pour le premier exemple 5.3.2, la figure 5.16 représente les moyennes $\bar{E}^{r_1}$ et $\bar{D}^{r_1}$, sur les 300 versions de l'algorithme, de l'erreur a posteriori E^{r_1} et de la distorsion D^{r_1}. Ces erreurs sont calculées sur des tableaux de données auxquels on a enlevé des composantes selon le procédé (1). La figure 5.17 représente les résultats analogues dans le cas (2).

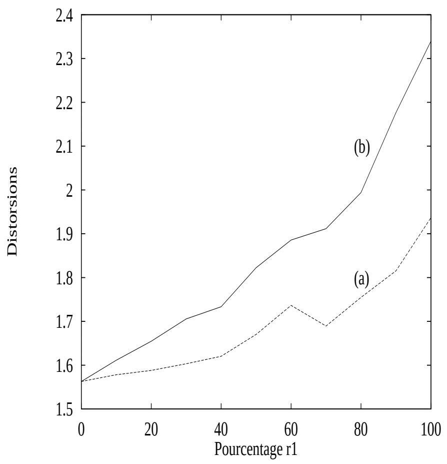


FIG. 5.16 - *Cas de trois classes : (a) Distorsion $\bar{D}^{r_1}$, (b) Erreur a posteriori $\bar{E}^{r_1}$.*

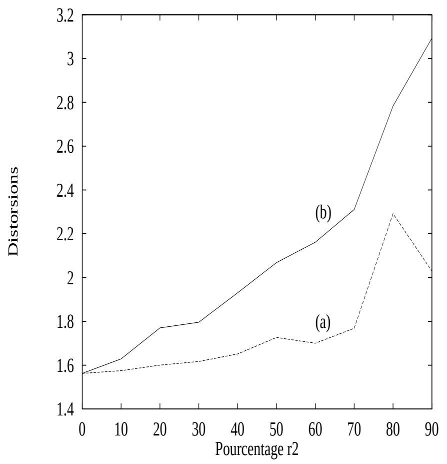


FIG. 5.17 - *Cas de trois classes : (a) Distorsion $\bar{D}^{r_2}$, (b) Erreur a posteriori $\bar{E}^{r_2}$.*

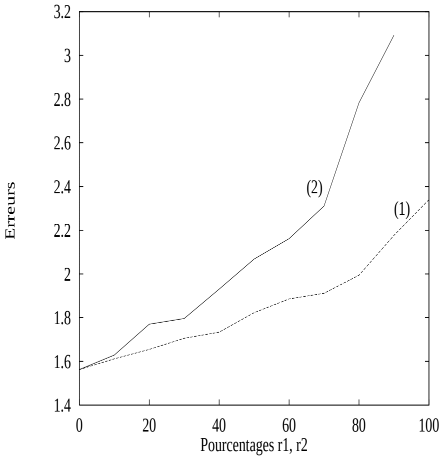


FIG. 5.18 - *Cas de trois classes : (1) Erreur $\bar{E}^{r_1}$; (2) Erreur $\bar{E}^{r_2}$.*

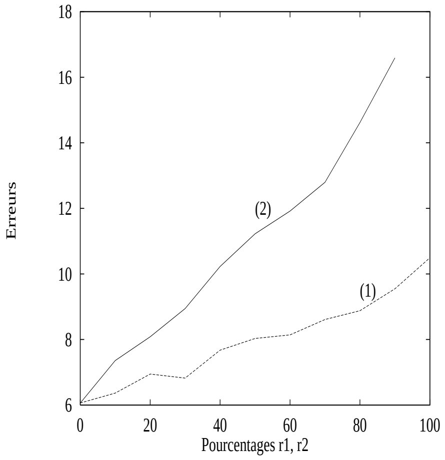


FIG. 5.19 - *Cas de quatre classes : (1) Erreur $\bar{E}^{r_1}$; (2) Erreur $\bar{E}^{r_2}$*

On peut remarquer que le fait de considérer l'erreur a posteriori E^r reste plus sévère que le fait de considérer la distorsion D^r. La figure 5.18 montre les erreurs résultant des deux considérations, c'est-à-dire des cas (1) et (2). D'où l'on peut constater que le cas (1) qui consiste à enlever m composantes choisies aléatoirement, est moins pénalisant que la situation (2) où l'on retire à chaque fois les mêmes composantes.

Cas de quatre classes

Sur cet exemple constitué de quatre classes (5.3.2), on retrouve des résultats semblables aux précédents.

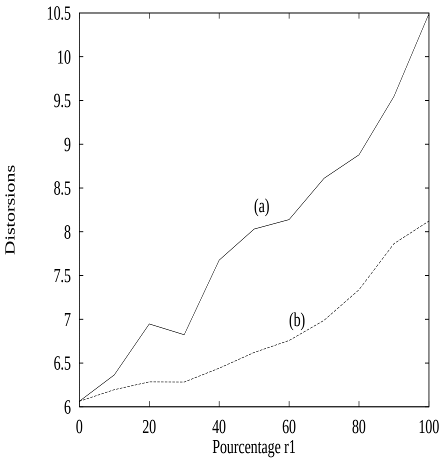


FIG. 5.20 - *Cas de quatre classes : (a) Erreur a posteriori $\bar{E}^{r_1}$, (b) Distorsion $\bar{D}^{r_1}$.*

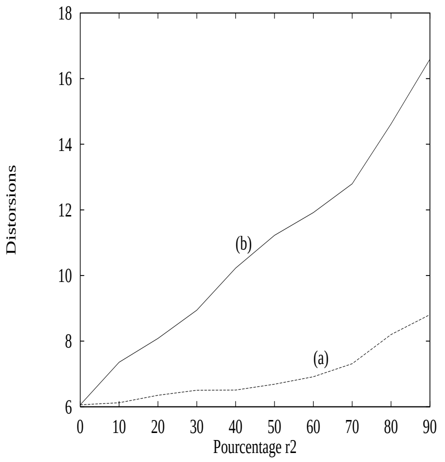


FIG. 5.21 - *Cas de quatre classes : (b) Erreur a posteriori $\bar{E}^{r_2}$, (a) Distorsion $\bar{D}^{r_2}$.*

On remarque donc le même comportement dans cet exemple. En effet dans la figure 5.19, l'erreur a posteriori dans le cas (2) est plus grande que celle dans le cas (1). De même les distorsions sont plus faibles que les erreurs a posteriori.

En conclusion, nous retenons que pour une proportion de valeurs manquantes donnée, il vaut mieux que ces valeurs non observées soient réparties dans toutes la base que concentrées sur quelques variables. Ainsi pour améliorer un résultat selon l'apprentissage de Kohonen en supprimant dans l'étape d'apprentissage des observations manquantes, il est préférable de raisonner en termes de variables plutôt que d'individus, c'est-à-dire d'enlever d'abord les variables qui ont « trop » de valeurs manquantes.

5.4 Estimation des valeurs manquantes

On suppose qu'on a fait un apprentissage d'un réseau de Kohonen grille ou ficelle avec une base de données incomplètes Ξ et on souhaite estimer les valeurs manquantes des individus.
La manière la plus simple pour estimer une valeur manquante est de remplacer celle-ci par la valeur correspondante du vecteur code auquel elle est associée. Soit

$$\forall x_j \in \Xi, \quad \forall k \notin M_{x_j}, \quad \hat{x}_j[k] = y_T^i[k]$$

$$\text{si } x_j \text{ est classé dans la classe de l'unité } i$$

où y_T^i est la valeur finale (après apprentissage du réseau) du vecteur code de l'unité i. Naturellement, pour les autres composantes qui sont présentes, on les conserve.

On a vu que lorsque l'on exécute plusieurs versions de l'algorithme de Kohonen, la distorsion et l'erreur a posteriori fluctuent, en fonction des versions, autour d'une valeur moyenne constante. On peut donc supposer que les valeurs des vecteurs poids eux aussi fluctuent autour d'une valeur moyenne. Ainsi, si on a V versions de cartes de Kohonen d'un même ensemble Ξ, alors on peut estimer cette fois ci les valeurs manquantes par

$$\forall x_j \in \Xi, \quad \forall k \notin M_{x_j}, \quad \hat{x}_j[k] = \frac{1}{V}\sum_v y_{T,v}^i[k]$$

où $y_{T,v}^i$ est la valeur finale du vecteur code i correspondant à la version v.

On note $\hat{X}$ le vecteur aléatoire d-dimensionnel qui est l'estimation (le complété) de X^p. Cette estimation étant faite par une des deux méthodes ci-dessus, c'est-à-dire à partir d'une version unique ou d'une moyenne sur plusieurs versions. On définit alors l'erreur d'estimation E_s par :

$$E_s = \mathbf{E}(\|X - \hat{X}\|) = \frac{1}{n}\sum_{j=1}^{n} \|x_j - \hat{x}_j\|$$

Dans le cas d'une estimation à partir d'une version unique, on a

$$E_s = \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{k \notin M_{x_j}} (x_j[k] - y_T^i[k])^2 \right)^{\frac{1}{2}}.$$

Car les autres termes de la deuxième somme sont nuls, puisque lorsque les composantes d'un individu sont présentes alors elles sont conservées dans $\hat{X}$.
On a alors,

$$E_s \leq E^p$$

L'erreur sur l'estimation E_s est une partie de l'erreur a posteriori, elle croît vers la borne supérieure quand p croît. L'allure de la courbe de E_s ressemble généralement à celle de E^p.

Comme l'algorithme de Kohonen finit l'apprentissage du réseau à « zéro voisin », le vecteur code d'une unité est par conséquent proche de la moyenne des individus qui sont affectés à cette unité (il faut supposer qu'à ce stade de l'apprentissage il n'y a pas trop de changement d'unités pour les vecteurs d'entrées). On peut en fait estimer les valeurs manquantes par la moyenne dans chaque unité. Il est évident que c'est beaucoup plus précis que d'estimer ces valeurs par la moyenne globale de la variable sur tous les individus. C'est donc une méthode naturelle, qui donne de bons résultats comme on va le vérifier dans les sections suivantes.

5.4.1 Les simulations

Nous reprenons les trois exemples précédents, pour voir les erreurs sur l'estimation dans les trois cas que nous avons distingués dans les sections précédentes. Nous regardons pour le cas des données socio-économiques d'Ile de France les erreurs commises sur quelques qu'on a choisies.

Les courbes ci-dessous représentent, pour les deux fichiers des exemples 4.11 et 5.3.2, les moyennes, sur 300 versions de cartes de Kohonen 8×8, des erreurs d'estimation en fonction de la proportion p, de valeurs manquantes, opérée sur chaque ligne de la table.

La figure 5.22 correspond à l'exemple 1 (trois classes), et la figure 5.23 correspond à l'exemple 2 (quatre classes). Dans les deux cas, la croissance de l'erreur est relativement faible jusqu'à $p = 50\%$. Il faudrait également regarder les erreurs relatives en pourcentages pour avoir une meilleure idée de l'erreur commise.

Les deux figures 5.24 et 5.25 correspondent aux cas où l'on enlève, au hasard, 7 composantes sur dix à une proportion r_1 croissante de lignes. Le cas où l'on enlève les sept premières composantes à une proportion r_2 croissante d'individus est représentée par les deux figures 5.26 et 5.27.

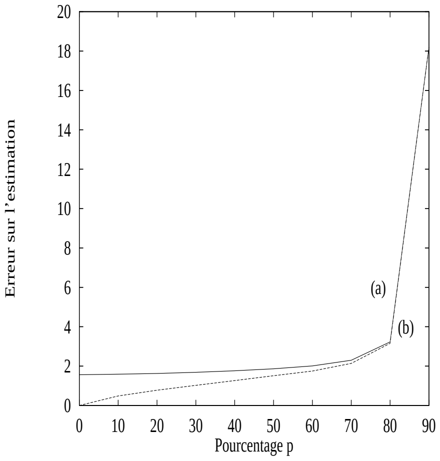


FIG. 5.22 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$.*

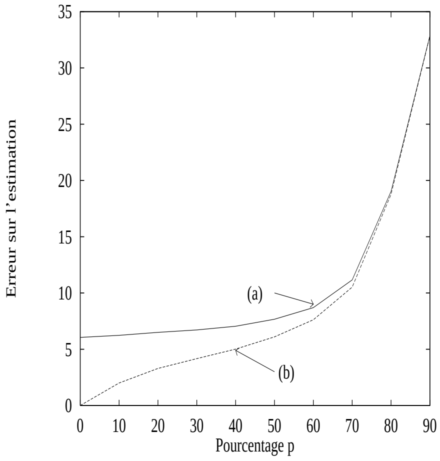


FIG. 5.23 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$.*

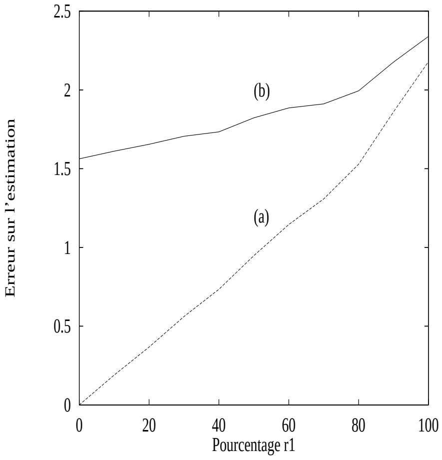


FIG. 5.24 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$; exemple 1.*

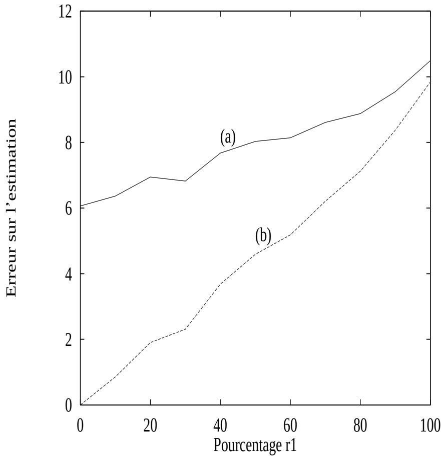


FIG. 5.25 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$; exemple 2.*

Il est clair, en regardant l'allure de ces courbes d'erreurs dans ces deux exemples, que l'erreur d'estimation croît plus rapidement en fonction de la proportion r_2 que de la proportion r_1.

Pour l'exemple des variables économiques de la base de données d'Ile de France, nous testons la robustesse et l'estimation par la carte de Kohonen sur trois tailles de réseaux : 8×8, 10×10 et 15×15.

La figure 5.28 montre qu'en dessous de $p = 60\%$ les courbes de l'erreur a posteriori decroissent en fonction de la taille du réseau (ce qui est logique), par contre pour $p \geq 60\%$, c'est le schéma inverse qui se produit. Il en ressort que lorsqu'il y a trop de valeurs manquantes il est préférable de ne pas choisir un réseau de grande taille.

Bien que la moyenne de l'erreur a posteriori $\bar{E}^p$ diminue en fonction de la taille du réseau, entre $p = 0\%$ et $p = 60\%$, l'erreur sur l'estimation ne semble pas être sensible à la taille

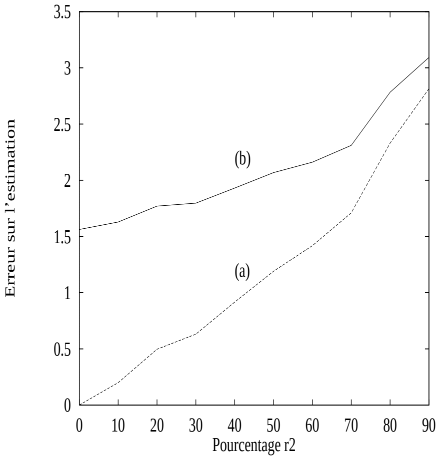


FIG. 5.26 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$; exemple 1.*

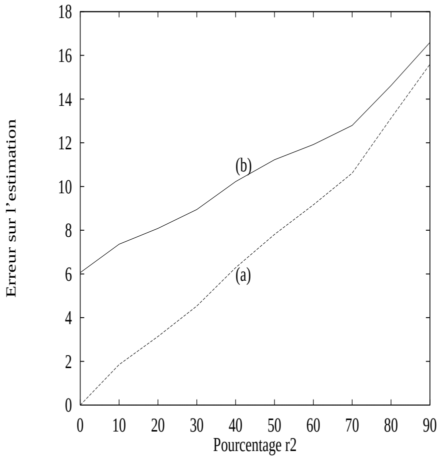


FIG. 5.27 - *(a) Erreur a posteriori $\bar{E}^p$, (b) Erreur sur l'estimation $\bar{E}_s$; exemple 2.*

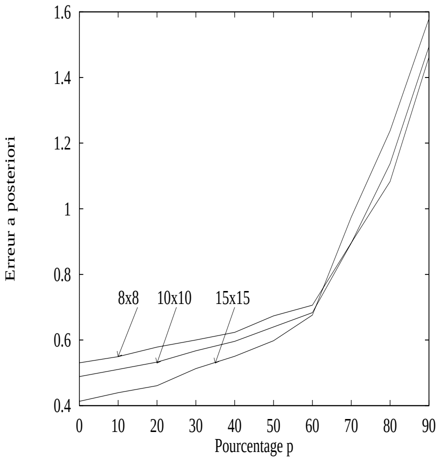


FIG. 5.28 - *Erreur a posteriori.*

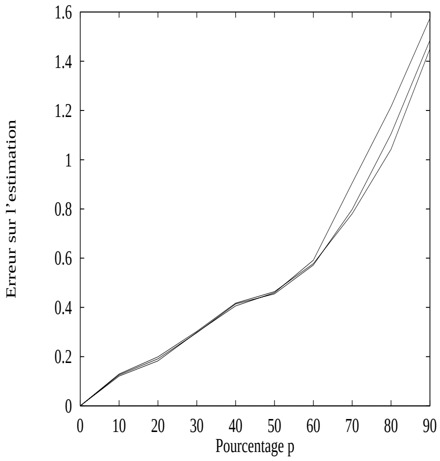


FIG. 5.29 - *Erreurs sur l'estimation .*

du réseau.

Le tableau 5.11 contient les moyennes sur 300 versions des estimations d'un exemple de base incomplète générée à partir du fichier sur les valeurs immobilières en Ile de France. Il s'agit d'un cas semblable à celui décrit en 5.3.3. On enlève à 40% d'individus les 7 premières composantes sur les 20 que contient la base. On compare les valeurs estimées aux valeurs réelles. On a choisi dix communes au hasard.

Lorsque l'on regarde les pourcentages d'erreurs, on a l'impression que la méthode n'est pas tout à fait adaptée. Mais il faut tenir compte du fait que la base de données est composée de 20 variables et que parmi ces 20 variables, certaines sont plus discriminantes que d'autres. Quand on regarde les cartes de Kohonen, on se rend compte que l'algorithme de Kohonen tient compte de ce pouvoir discriminant. Comme par exemple, dans l'étude des valeurs immobilières en Ile de France, la dimension des prix est très forte (Cf 4.3.3). Il arrive donc qu'une variable ne soit pas bien représentée sur une carte, c'est le cas visiblement de la variable six. Par contre, pour la variable deux, hormis les communes (ARPA, AVON et pa18), la variable est assez bien estimée. Il y a en plus visiblement des mauvais classements tels que pour les communes (ARPA et AVON) pour lesquelles les estimations sont mauvaises partout. On remarque par contre que pour les classes homogènes telles que Paris (pa16, pa17 et pa18), les estimations ne sont pas mauvaises. Il faut donc tenir compte de ces remarques pour l'estimation des valeurs manquantes.

commune	Valeurs réelles						
	variable 1	variable 2	variable 3	variable 4	variable 5	variable 6	variable 7
5ANTO	0.671	0.694	0.220	0.098	0.379	1.583	0.161
2ARPA	0.668	0.445	0.102	0.119	0.310	0.345	0.310
1ASNI	0.730	0.564	0.167	0.163	0.361	0.866	0.218
3ATHI	0.547	0.544	0.099	0.124	0.288	0.402	0.271
1AVON	0.765	0.602	0.171	0.103	0.355	0.792	0.216
2BEZO	0.434	0.532	0.058	0.185	0.249	0.183	0.346
1BGNX	0.608	0.592	0.123	0.121	0.315	0.535	0.239
6pa16	1.462	0.877	0.400	0.164	0.420	5.346	0.077
6pa17	1.200	0.751	0.320	0.143	0.434	2.478	0.129
6pa18	1.047	0.635	0.201	0.200	0.391	1.035	0.202
commune	Valeurs estimées						
5ANTO	0.658	0.680	0.242	0.078	0.399	2.480	0.129
2ARPA	0.471	0.567	0.095	0.131	0.277	0.423	0.282
1ASNI	0.874	0.588	0.173	0.124	0.384	1.133	0.196
3ATHI	0.484	0.575	0.100	0.122	0.286	0.469	0.268
1AVON	0.688	0.688	0.263	0.085	0.398	2.980	0.123
2BEZO	0.438	0.555	0.080	0.154	0.253	0.299	0.312
1BGNX	0.465	0.553	0.090	0.173	0.254	0.354	0.305
6pa16	1.237	0.828	0.375	0.138	0.451	4.847	0.097
6pa17	1.154	0.753	0.312	0.154	0.437	2.931	0.134
6pa18	1.107	0.741	0.298	0.164	0.427	2.669	0.146
commune	Erreurs en pour cent						
5ANTO	2	2	10	20	5.2	56	20
2ARPA	30	27	7	10	10	22	9
1ASNI	19	4	3.5	24	6	30	10
3ATHI	11	6	1	1.6	0.5	16	1
1AVON	10	14	53	18	12	270	43
2BEZO	0.2	4	27	17	2	60	10
1BGNX	23	7	27	40	19	33	27
6pa16	15	6	6	15	7	9	28
6pa17	4	0.2	2.5	7.6	0.7	14	4
6pa18	5	16	32	18	9	70	27

TAB. 5.11 - *Tableau d'erreurs sur les estimations en pour cent*

5.5 Conclusion

Sur le traitement des données manquantes, on peut faire deux types de conclusions : d'une part sur la classification des individus et d'autre part sur l'estimation des valeurs manquantes.

D'abord la méthode KACP pour les valeurs manquantes est naturelle et facile à mettre en oeuvre. Du point de vue de la classification des individus, la méthode parait bien robuste. Le classement des individus se conserve jusqu'à un pourcentage relativement grand de valeurs manquantes. De plus si on a des super classes bien séparées, alors elles restent toujours bien séparées jusqu'à soixante dix pour cent de valeurs manquantes par individu.

Sur l'estimation par les vecteurs codes du réseau de Kohonen, la méthode semble être très sensible aux mauvais classements dus aux valeurs manquantes ou à l'héterogénéité de la base elle même. Il faut donc faire des estimations sur des classes bien homogènes et en utilisant des variables de préférence bien corrélées entre elles.

Chapitre 6

La méthode KACM pour les variables qualitatives

6.1 Introduction

Ce chapitre est consacré à l'utilisation de l'algorithme de Kohonen pour le traitement de variables qualitatives. Nous utilisons dans ce chapitre les notations et les résultats du chapitre 3. On sait que dans le cas des variables quantitatives, l'algorithme de Kohonen fait une quantification vectorielle de l'espace des individus par rapport aux observations de d variables continues. On a, par cette méthode, une classification et une carte organisée semblable à une projection de type ACP (Cf 4).

On rappelle (Cf 3.3 et 3.4) que dans le cas des variables qualitatives, on cherche à mettre en évidence la typologie des modalités et on essaye de faire ressortir les relations qui existent entre les modalités des différentes variables. Par exemple, s'il y a une proportion non négligeable d'individus qui ont choisi la modalité «a» de la question «1» et la modalité «b» de la question «2», alors on dira que les modalités (1,a) et (2,b) sont liées, et qu'elles s'attirent. Si par contre il y a une proportion importante d'individus qui choisissent (1,a) et qui rejettent (2,b), alors elles se repoussent.

On voit bien que l'algorithme de Kohonen qui organise les unités en respactant les proximités dans l'espace des entrées peut tout à fait convenir à ce type de traitement. La première méthode qui a traité les données qualitatives, au moyen de l'algorithme de Kohonen, date de 1993 et est due à (Cottrell et al., 1993). Baptisée « *Kouplet* », cette méthode concerne le cas où le nombre de variables (ou questions) Q, est égal à deux. En effet, on utilise les profils lignes et les profils colonnes d'un tableau de contingence pour produire deux cartes de Kohonen; la première concerne les profils lignes et la seconde les profils colonnes. On procède à un « couplage » des deux variables en associant à chaque profil ligne le profil colonne qui lui est le plus probable, c'est-à-dire celui avec lequel il a la plus grande probabilité conjointe, on obtient ainsi deux cartes couplées qu'on peut superposer. La méthode donne de bons résultats et a été utilisée dans plusieurs cas concrets.

Nous avons proposé, en 1995 dans (Ibbou and Cottrell, 1995b), une méthode baptisée « KACM », qui permet de classer les modalités de Q variables qualitatives, où Q est supérieur ou égal à deux. Nous améliorons, dans ce chapitre, cette méthode en la simplifiant et indiquons également comment « projeter » simultanément les individus et les modalités des variables sur une carte de Kohonen. Ce deuxième apport a un intérêt particulier, puisque en réalisant cette carte de Kohonen qui classe simultanément les individus et les modalités des variables, on obtient non seulement une association des modalités, mais aussi une classification des individus par rapport à ces modalités. Nous verrons que cette projection peut se faire de deux manières différentes, soit au moyen du tableau disjonctif complet, soit au moyen du tableau de Burt. La deuxième technique est plus économique du point de vue du temps de calcul, et ceci principalement dans le cas d'une base de données contenant un nombre d'individus très important par rapport au nombre de modalités.

6.2 Rappels

On rappelle les notations utilisées dans le chapitre 3. On suppose qu'on a les réponses de N individus à Q variables qualitatives, avec $Q \geq 2$. Chaque question q, $1 \leq q \leq Q$, possède m_q modalités et on a M modalités au total. Les tableaux d'entrées peuvent être : le tableau disjonctif complet, la table de Burt ou bien dans le cas où $Q = 2$, un tableau de contingence.

On note $K_{(N \times M)}$, la matrice du tableau disjonctif complet :

$$K = (k_{ij}) \quad \text{avec} \quad k_{ij} = \begin{cases} k_{ij} = 1 & \text{si l'individu } i \text{ a choisi la modalité } j \\ k_{ij} = 0 & \text{sinon} \end{cases}$$

et on a $\forall i, 1 \leq i \leq N, \quad k_{i.} = \sum_{j=1}^{j=M} k_{ij} = Q$. La marginale sur les lignes est donc constante égale au nombre de questions. On note W la matrice diagonale constituée par les éléments de la marginale sur les colonnes.

$$W = Diag\{k_{\cdot 1}, \ldots, k_{\cdot i}, \ldots, k_{\cdot M}\}.$$

Le tableau K n'est pas toujours disponible; on peut résumer l'information sur les modalités des variables (en perdant une partie de l'information sur les individus) dans une table de Burt $B_{(M \times M)}$ définie par

$$B = (b_{ij}) = K^t K.$$

C'est une matrice symétrique carrée telle que :

$$\forall i, \quad 1 \leq i \leq M, \quad b_{i.} = b_{.i} = Q k_{.i}.$$

Dans ce chapitre, nous n'utilisons que ces deux tableaux. Dans le cas où $Q = 2$, si l'on dispose uniquement du tableau de contingence $C = (c_{ij})$, $1 \leq i \leq m_1$ et $1 \leq j \leq m_2$, alors on se ramène à un tableau de Burt par

$$B = \begin{pmatrix} C_{i\cdot} & C \\ C^t & C_{\cdot j} \end{pmatrix}.$$

où

$$C_{i\cdot} = diag\{c_{1\cdot}, ..., c_{m_1\cdot}\} \quad C_{\cdot j} = diag\{c_{\cdot 1}, ..., c_{\cdot m_2}\}$$

sont des matrices diagonales constituées respectivement de la marginale sur les lignes et de la marginale sur les colonnes.

Nous avons établi dans la section (3.4), en utilisant la matrice K correspondant au tableau disjonctif complet, que les axes principaux maximisant l'inertie projetée des individus sont portés par les vecteurs propres de la matrice

$$\frac{1}{Q}K^tKW^{-1} = \frac{1}{Q}BW^{-1}.$$

D'autre part en utilisant la matrice de Burt, on sait que les axes maximisant l'inertie projetée des profils lignes sont portés par les vecteurs propres de la matrice

$$\frac{1}{Q^2}B^tW^{-1}B^tW^{-1}.$$

On rappelle brièvement (Cf 3.4) que l'ACM équivaut à une ACP pondérée sur les profils lignes ou colonnes, munis d'une métrique particulière qui est la métrique du Chi-deux.

En effet, pour une ACM à partir du tableau disjonctif complet (Cf 3.4.1), on considère la matrice des profils lignes correspondant au tableau K, (c'est la matrice $\frac{1}{Q}K$), sur laquelle on effectue une ACP, en utilisant la métrique définie par la matrice NQW^{-1} et la pondération des profils lignes définie par la matrice $\frac{1}{N}I_N$.

Dans le cas de l'ACM à partir de la matrice de Burt (Cf 3.4.2), on utilise cette fois-ci la matrice des profils lignes $\frac{1}{Q}W^{-1}B$, la métrique définie par la matrice QNW^{-1} et la matrice de pondération définie par QNW^{-1}. Ceci nous conduit à utiliser un algorithme de Kohonen, adapté à chacun de ces deux cas.

Les entrées sont les vecteurs lignes des matrices des profils lignes, et on choisit la même métrique, et la même pondération qu'en ACM. En fait, il est beaucoup plus commode de se ramener à la pondération uniforme et à la distance euclidienne, car on peut ainsi utiliser l'algorithme standard.

De plus, bien que les démonstrations théoriques soient faites pour une distribution quelconque des entrées et pour n'importe quelle distance, les paramètres adaptatifs, généralement empiriques, sont le plus souvent testés et maîtrisés dans le cas de la loi uniforme et de la distance euclidienne usuelle. C'est pour cette raison, et pour pouvoir utiliser un algorithme de Kohonen standard, qu'on essaye de se ramener à ce cas, en utilisant une matrice de Burt corrigée ou un tableau disjonctif complet corrigé comme expliqué ci-dessous.

6.3 Classification des Modalités

Il faut tout d'abord choisir l'ensemble des vecteurs d'entrée. Comme ces vecteurs d'entrée sont généralement les lignes d'un tableau de données, nous utilisons la matrice de Burt corrigée B^c introduite à la page 77.

$$B^c = (b^c_{ij}) \quad \text{avec} \quad b^c_{ij} = \frac{b_{ij}}{\sqrt{b_{i\cdot}}\sqrt{b_{\cdot j}}}$$

dont les lignes seront les vecteurs d'entrée de l'apprentissage du réseau de Kohonen. Comme nous l'avons vu, grâce à cette transformation, on peut choisir les entrées selon une loi uniforme et utiliser la distance euclidienne au cours de l'apprentissage. On prend un réseau de Kohonen carré ou un réseau ficelle si l'on pense que les modalités sont ordinales. Chaque unité, u du réseau R, est munie d'un vecteur code y^u_t de $\mathbb{R}^M$. On peut alors définir l'algorithme KACM comme suit.

l'algorithme KACM

- A chaque itération on choisit, uniformément parmi toutes les lignes de la matrice B^c, une ligne b^c_l :

$$b^c_l = (\frac{b_{l1}}{\sqrt{b_{l\cdot}}\sqrt{b_{\cdot 1}}}, \ldots\ldots, \frac{b_{lM}}{\sqrt{b_{l\cdot}}\sqrt{b_{\cdot M}}}).$$

- On cherche l'unité u_0 gagnante parmi l'ensemble des unités du réseau : c'est l'unité qui réalise le minimum de l'écart quadratique usuel.

$$u_0 = \arg\min_{u \in R} \|y^u_t - b^c_l\|^2$$

- On actualise les vecteurs codes de l'unité u_0 et de ses unités voisines par

$$y^u_{t+1} = y^u_t + \epsilon(t)(b^c_l - y^u_t)$$

Après apprentissage, on classe les modalités (représentées par les lignes de B^c) dans le réseau en affectant chacune d'elles à l'unité u du réseau qui lui est la plus proche au sens des moindres carrés.

Généralement le nombre de modalités n'est pas très grand; l'apprentissage du réseau est par conséquent très rapide. Nous présentons les cartes produites par cette méthode originale sur différents exemples.

6.3.1 Exemples

Nous appliquons cette méthode KACM à trois exemples scolaires; deux sont pris dans la bibliothèque d'exemples du logiciel SAS, l'exemple sur les races canines et leurs caractéristiques est tiré du livre de (Saporta, 1990). D'autres cas réels ont été traités par cette méthode dont une étude sur les entreprises belges pratiquant le leasing et une étude pour le ministère des transports sur les moyens utilisés par les voyageurs en fonction des distances parcourues.

6.3.2 Les caractéristiques physiques

Ce premier exemple est tiré des exemples que le logiciel SAS présente dans la procédure « Corresp ». Il est simple et instructif car il ne contient que 11 individus et 4 variables qui sont :

- **Age** en deux modalités (Vieux,Jeune)
- **Sexe** en deux modalités (Homme,Femme)
- **Taille** en deux modalités (Grand,Petit)
- **Cheveux** en trois modalités (Blanc,Brun,Blond)

Nom	âge		sexe		taille		cheveux		
	Vieux	Jeune	Home	Femme	Petit	Grand	Blanc	Brun	Blond
jones	1	0	1	0	1	0	1	0	0
smith	0	1	0	1	0	1	0	1	0
kesavitz	1	0	1	0	1	0	0	1	0
ernst	1	0	0	1	0	1	1	0	0
zannoria	1	0	0	1	1	0	0	1	0
spangel	0	1	1	0	0	1	0	0	1
myers	0	1	1	0	0	1	0	1	0
kasinski	1	0	1	0	1	0	0	0	1
colman	0	1	0	1	1	0	0	0	1
delafave	1	0	1	0	0	1	0	1	0
singer	0	1	1	0	0	1	0	1	0

TAB. 6.1 - *Tableau disjonctif complet.*

A partir des lignes de la matrice de Burt corrigée correspondant au tableau disjonctif complet (6.1), on fait l'apprentissage d'un réseau de Kohonen en grille et en ficelle. Les cartes ci-dessous représentent les résultats obtenus après 200 itérations.

Compte tenu du fait que cet exemple n'est pas très riche en individus, l'algorithme de Kohonen peut avoir quelques difficultés à former des groupes de modalités stables durant

Brun Grand	Jeune	Homme	
			Blond
Blanc Vieux	Petit		Femme

TAB. 6.2 - *Une grille à 4×4 unités*

Brun Grand		Jeune	Blond	
	Homme			Femme
		Blanc	Vieux	Petit

TAB. 6.3 - *Une grille à 5×5 unités.*

Vieux Blanc		Femme Petit	Blond		Homme Jeune		Brun Grand

TAB. 6.4 - *Une ficelle à 8 unités.*

les versions successives. On remarque cependant que dans les trois résultats obtenus ci-dessus (6.2, 6.3 et 6.4) on retrouve des associations similaires qui sont (Grand, Brun) et (Vieux,Blanc).

Pour vérifier ces résultats, on compare les cartes de Kohonen aux projections obtenus par une ACM classique. Les figures (6.1, 6.2) représentent les projections sur les trois premiers axes principaux.

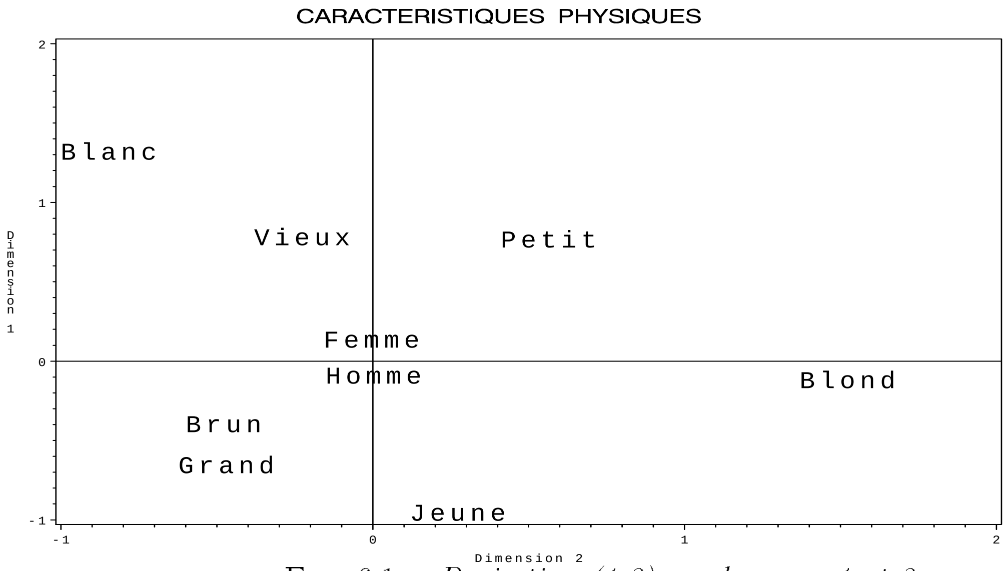


FIG. 6.1 - *Projection (1,2) sur les axes 1 et 2*

Sur la projection (1,2) on retrouve bien les associations, annoncées par les cartes de Kohonen, puisqu'on a Brun à coté de Grand et Jeune d'une part et Blanc à proximité de Vieux. Mais on a également la modalité Femme qui se projette à proximité de la modalité Homme, alors que ce n'est pas le cas dans la projection (1,3).

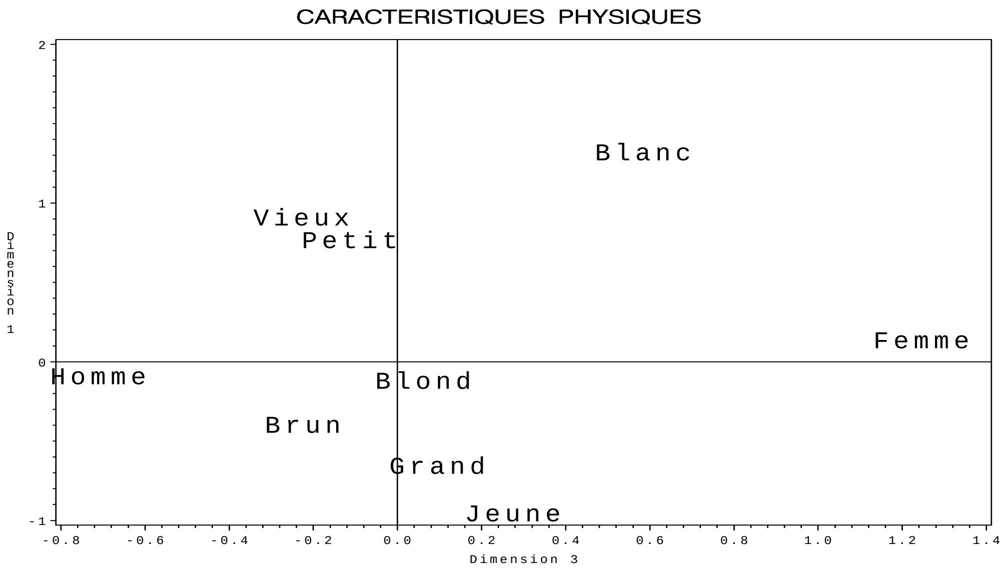


FIG. 6.2 - *Projection sur les axes 1 et 3*

La représentation de Kohonen permet d'éviter de « fausses » proximités comme celles « Homme-Femme » dans la carte (1,2). L'ACM exige la consultation de plusieurs cartes successives, alors que la représentation de Kohonen est unique.

6.3.3 Voitures et leurs propriétaires

Cet exemple est également extrait de ceux que le logiciel SAS version 6.10 propose pour illustrer la procédure « Corresp ». On a le tableau de Burt qui croise 7 variables :

- **Origine de la voiture** (Américaine, Japonaise, Européenne)
- **Taille de la voiture** (Petite, Moyenne, Grande)
- **Type de voiture** (Familiale, Sport, Travail)
- **Type de possession** (Propriétaire, Location)
- **Statut familial** (Célibataire, Mariés, Célibataire et enfants, Mariés et enfants)
- **Revenu** (1-revenu, 2-revenus)
- **sexe** (Homme, Femme)

On a alors $M = 19$, et le résultat de l'algorithme KACM, sur une grille 5×5, figure dans le tableau 6.5 ci-dessous

1-revenu **Célibataires**	Location			**Célibat&Enf**
	Sport		*Européenne*	Travail
Japonaise Homme PETITE				GRANDE
		Femme		*Américaine*
2-revenus **Mariés**		Propriétaire	Moyenne	**Mariés&Enf** FAMILIALE

TAB. 6.5 - *On remarque aux quatre coins de la carte que les Célibataires sont opposés aux Mariés et les ménages avec enfants sont opposés aux ménages sans enfants.*

On observe les regroupements de (1-revenu et célibataires), (2-revenus et mariés), (Familiale et Mariés avec Enfants). De plus les deux unités aux extrémités supérieures représentant les célibataires sont opposées aux extrémités du bas représentant les Mariés.

L'analyse des correspondances multiples classique donne des résultats similaires sur les projections dans le plan (1,2) figure 6.3 et dans le plan (1,3) figure 6.4.

Nous constatons les mêmes grandes relations : association entre 2-revenus et mariés, voiture familiale avec mariés avec enfant, célibataire avec 1-revenu et location, petite voiture avec japonaise etc...

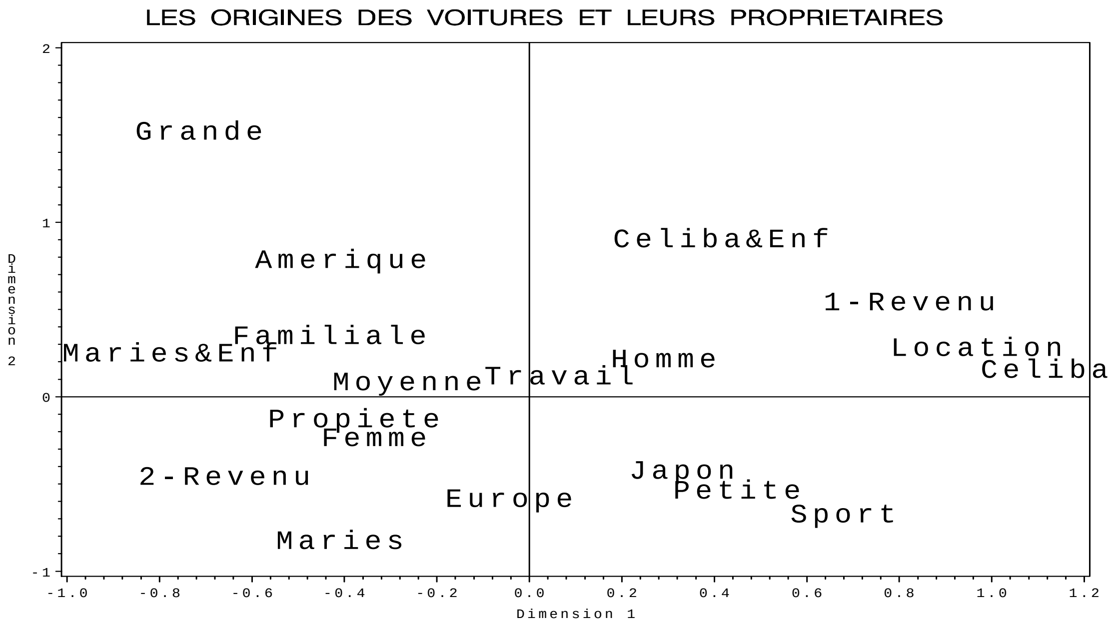


FIG. 6.3 - *Projection sur les axes 1 et 2*

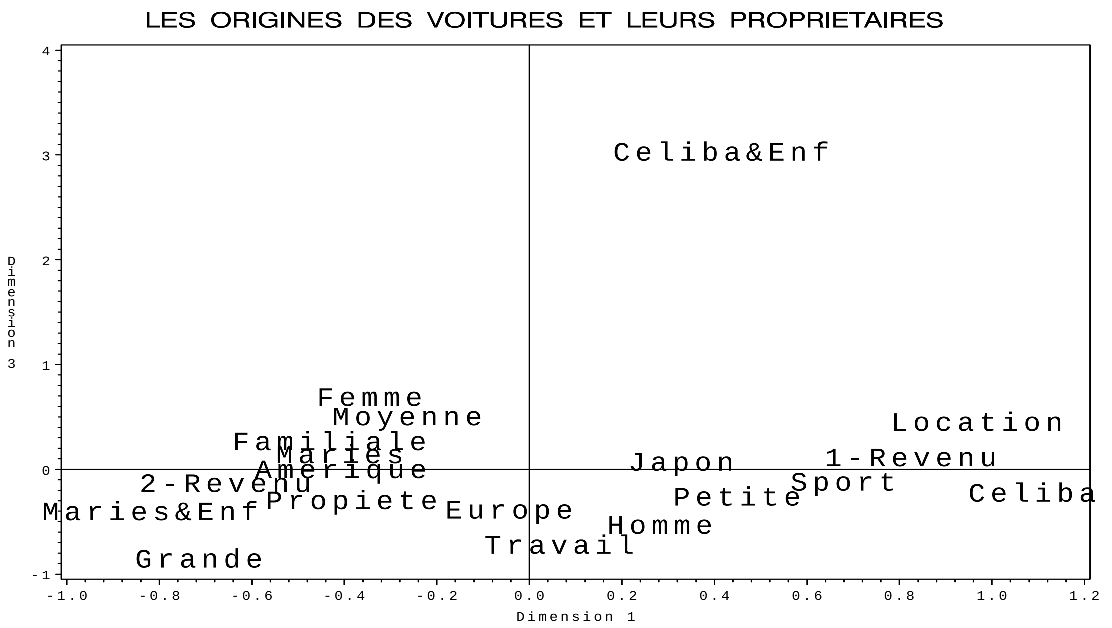


FIG. 6.4 - *Projection sur les axes 1 et 3*

6.3.4 Les races de chiens

Cet exemple est tiré du livre de (Saporta, 1990). On regarde les réponses que donnent 27 races de chiens sur les 7 variables suivantes :

- **Vélocité** en trois modalités (Petite,Moyenne,Grande),
- **Taille** en trois modalités (Petite, Moyenne, Grande),
- **Poids** en trois modalités (Petite, Moyenne, Grande),
- **Affection** en deux modalités (Affectueux, Non affectueux),
- **Intelligence** en trois modalités (Petite, Moyenne, Grande),
- **Agressivité** en deux modalités (Agressif, Non agressif),
- **Fonction** en trois modalités (Utile, Chasse, Compagnie).

Le tableau disjonctif complet ci-dessous 6.6 résume les réponses pour chaque race à chaque modalité.

Race	PtTaille	MyTaille	GdTaille	GdPoids	MyPoids	GdPoids	PtVéloc	MyVéloc	GdVéloc	PtIntel	MyIntel	GdIntel	NonAffe	Affect	Nagres	Agressif	Compagnie	Chasse	Utile
Beauceron	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1
Basset	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0
Berger Allemand	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1
Boxer	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	0
Bull-Dog	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
Bull-Mastiff	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	1
Caniche	1	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0
Chihuahua	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
Cocker	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
Colley	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	0
Dalmatien	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	0
Doberman	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1
Dogue Allemand	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1
Epagneul Breton	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0
Epagneul Français	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0
Fox-Hound	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	0
Fox-Terrier	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0
Grand Bleu de Gascogne	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0
Labrador	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0
Levrier	0	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0	0	1	0
Mastiff	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	1
Pekinois	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
Pointer	0	0	1	0	1	0	0	0	1	0	0	1	1	0	1	0	0	1	0
Saint-Bernard	0	0	1	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	1
Setter	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0
Teckel	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1	0	0
Terre-Neuve	0	0	1	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	1

TAB. 6.6 - *Les caractéristiques Canines, d'après le livre*

On calcule la matrice de Burt corrigé B^c et avec les lignes de cette matrice on procède à l'apprentissage d'un réseau carré constitué de 5×5 noeuds. Après 300 itérations on obtient la grille suivante :

PtPoids PtTaille	PtVéloc	PtIntel	Agressif	**Utile** *GdPoids*
Compagnie				
Affectueux		GdIntel		NAffect GdTaille
MyIntel	NAgressif			
MyVéloc MyTaille		**Chasse** *MyPoids*		GdVéloc

La carte est organisée et l'on obtient des groupes de modalités telles que (Compagnie, Petite taille, Petit poids), (Grande taille, Utile), (Moyenne taille, Chasse)...

On peut comparer avec la projection de ces 19 vecteurs sur les deux premières directions principales obtenus par l'ACM classique (réalisée au moyen du logiciel SAS); on obtient des relations similaires.

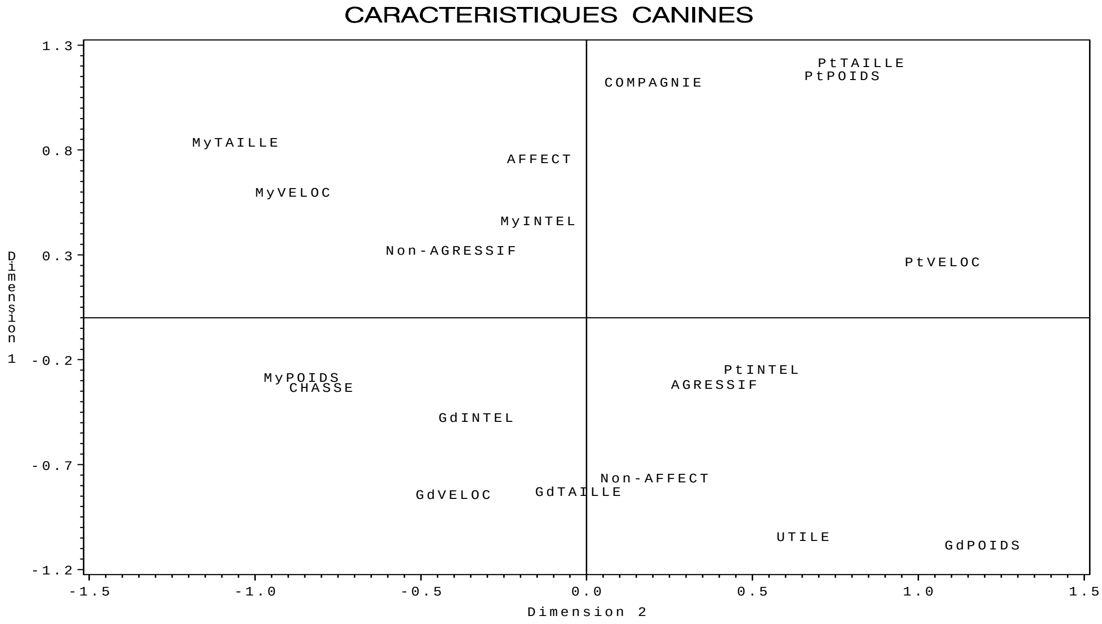


Fig. 6.5 - *Projection sur les axes 1 et 2*

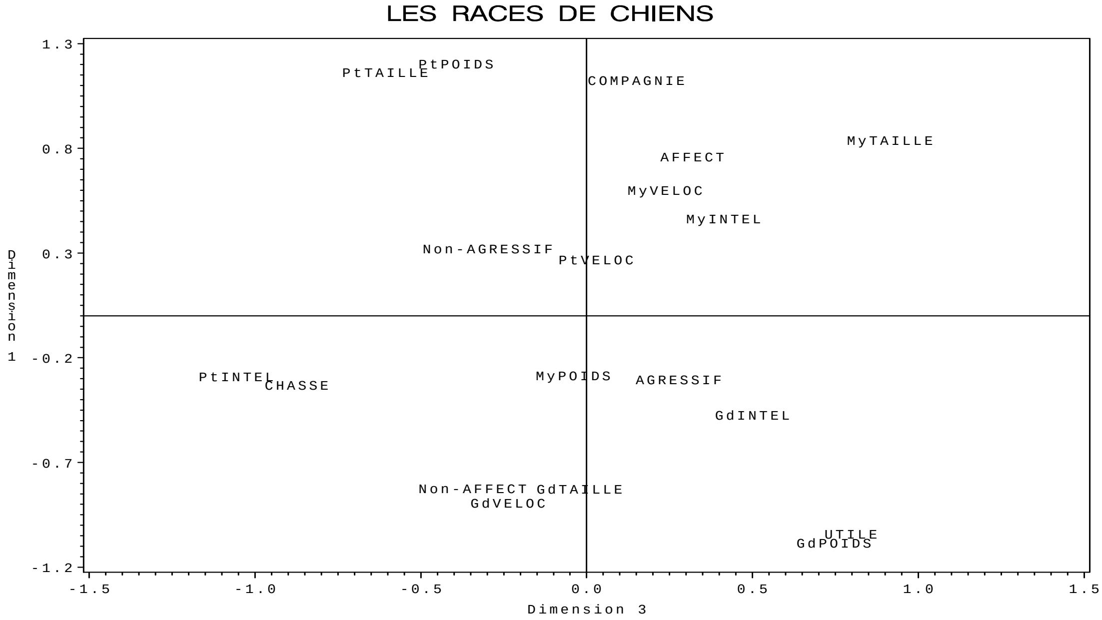


FIG. 6.6 - *Projection sur les axes 1 et 3*

On a bien les associations (Grand poids, Utile), (Petite taille, Petit poids et Compagnie) ect.

6.4 Classification des individus et des variables

Tout d'abord, pour pouvoir faire figurer les individus dans des cartes de Kohonen, il faut disposer du tableau disjonctif complet. Or nous avons déjà établi (page 77) que les axes principaux d'une ACM utilisant le tableau disjonctif sont les axes portés par les vecteurs propres de la matrice

$$I_K = \frac{1}{Q} K^t K W^{-1} = (K^c)^t K^c,$$

où K^c est définie par :

$$K^c = (k^c_{ij}) \quad \text{avec} \quad k^c_{ij} = \frac{k_{ij}}{\sqrt{k_{i\cdot}}\sqrt{k_{\cdot j}}}.$$

D'autre part, on sait que si on utilise la matrice de Burt on cherche les axes principaux de la matrice

$$I_B = \frac{1}{Q^2} B W^{-1} B W^{-1} = (B^c)^t B^c.$$

Nous avons déjà remarqué dans le chapitre 3 que si u est vecteur propre de I_K avec une valeur propre λ, alors u est vecteur propre de I_B associé à la valeur propre λ^2. Ce sont donc les mêmes axes qui maximisent l'inertie des deux nuages de points par rapport aux modalités. En réalité la projection simultanée des individus et des modalités pose un problème de cohérence, qui n'est pas éclairci totalement dans la littérature.

En effet dans le cas du tableau de Burt, on a une matrice $M \times M$ où les lignes et les colonnes représentent les modalités. Par contre, dans le cas du tableau disjonctif complet, il s'agit d'un tableau $N \times M$ où les lignes représentent les individus et les colonnes représentent les modalités. Dans les deux cas, les vecteurs colonnes représentent les mêmes modalités, mais dans un cas, ce sont des vecteurs de $\mathbb{R}^M$ et dans l'autre ce sont des vecteurs de $\mathbb{R}^N$. On sait que l'ACM classique construit des « modalités génériques » (composantes principales) qui sont des combinaisons linéaires des modalités originelles, c'est-à-dire des vecteurs colonnes. Mais étant donné que les dimensions sont distinctes, la superposition des nouveaux systèmes d'axes (composantes principales) est seulement approximative, qui se justifie en considérant qu'on superpose les variables canoniques les plus corrélées. Dans la méthode KACM, il n'y a pas de conflit de dimensions, car on considère les vecteurs lignes de ces deux matrices qui sont tous des vecteurs de $\mathbb{R}^M$.

Comme pour la méthode KACM décrite dans 6.3, on se ramène à la distance euclidienne et à la pondération uniforme en utilisant la matrice K^c et la matrice B^c pour classer simultanément les individus et les modalités. On distingue donc deux méthodes **KACM I** et **KACM II**; la première utilisant la matrice corrigée K^c pour l'apprentissage du réseau et la seconde la matrice B^c.

6.4.1 L'algorithme KACM I : Apprentissage par le Tableau Disjonctif Complet

On suppose que le choix du réseau est fait, le nombre d'unités et son architecture sont fixés. Les vecteurs codes des unités sont des vecteurs de $\mathbb{R}^M$ qu'on initialise aléatoirement. Avec les lignes de cette matrice K^c, on procède à l'apprentissage du réseau :

- A chaque itération on choisit, uniformément parmi toutes les lignes de la matrice K^c, une ligne $k_l^c = (\frac{k_{l1}}{\sqrt{k_{l\cdot}}\sqrt{k_{\cdot 1}}}, \ldots\ldots, \frac{k_{lM}}{\sqrt{k_{l\cdot}}\sqrt{k_{\cdot M}}})$.
- On cherche l'unité u_0 gagnante parmi l'ensemble des unités du réseau : c'est l'unité qui réalise le minimum de la distance quadratique usuelle.
- On actualise les vecteurs codes de l'unité u_0 et de ses unités voisines par la formule standard.

Après apprentissage, on classe chaque individu dans le réseau en l'affectant à la classe définie par son unité gagnante. On obtient alors une carte de Kohonen où seuls les individus sont classés.

Pour faire figurer les modalités sur la même carte de Kohonen, on procède comme suit : avec les lignes de la matrice B^c, on classe les modalités dans la carte en affectant chacune d'elles à l'unité du réseau qui lui est la plus proche. La distance qui est utilisée ici est la distance euclidienne usuelle.

Bien que l'apprentissage n'ait pas été fait avec les lignes correspondant aux modalités, le réseau obtenu peut servir pour classer les modalités, mais il n'est pas un bon quantificateur vectoriel pour les modalités. En fait les groupes d'individus qui se ressemblent pour avoir choisi les mêmes modalités vont attirer ces mêmes modalités. Les vecteurs codes des unités ne sont de bons représentants que pour les lignes de la matrice K^c, mais ils ne le sont pas pour les lignes de la matrice B^c.

Exemple des races canines

Les trois tableaux suivants sont les résultats de l'application de la méthode KACM I pour le tableau des caractéristiques canines. La carte représentée par le tableau 6.7 illustre une version de la méthode KACM I où l'on fait figurer uniquement les races de chiens sans les modalités.

On peut voir qu'on a les chiens de compagnie (Bull-Dog, Teckel, Pekinois...) qui se regroupent autour de noeuds voisins, les chiens de grande taille (Dogue Allemand, Bull Mastif, Saint Bernard...) qui se regroupent également autour des noeuds en haut à droite. On a également (Berger Allemand, Doberman) et (Pointer, lévrier) etc.

Pour en savoir plus sur les modalités qui caractérisent ces regroupements de races, on fait figurer également les modalités des variables sur la grille. On a dans la carte (tableau 6.8) les associations de (Chasse, Setter, Levrier, Pointer) et à proximité de (Poids Moyen). De plus, on a (Compagnie, Petit poids, Teckel, caniche...) à proximité de (Affectueux) et (Petite Taille). La ficelle obtenue par cette méthode (tableau 6.9) montre certaines des associations logiques telles que celles précédemment citées, mais on remarque qu'on a (Grande Intelligence et Petite Intelligence) dans la même unité (la deuxième en partant de la droite).

Beauceron Berge-All			Fox-Hound			Mastiff	Saint-Ber Terre-Neu
	Doberman				Dogue-All		Bull-Mast
			Levrier				
Basset			Pointer		Setter		GrBlGasco
						Epagneu-F	
Chihuahua Pekinois		Caniche		Colley		Labrador	Epagneu-B
Bull-Dog Teckel		Fox-Terri		Cocker		Boxer	Dalmatien

TAB. 6.7 - *Carte de Kohonen sur les individus; apprentissage avec la matrice K^c*

Compagnie Cocker		**Affect**		**MyIntell**			**MyTaille** Boxer Dalmatien Labrador
Fox-Terri				Non-Agre Colley		**MyVitesse**	Epagneu-B
PtPoids Bull-Dog Teckel	Caniche				Epagneu-F		
	PtTaille	Chihuahua Pekinois			**Chasse** Setter	**MyPoids**	
			Basset		Fox-Hound GrBlGasco Levrier	Pointer	**GrIntell** Beauceron Berge-All Doberman
				PtIntell			**GrVitesse**
	PtVitesse						**GrTaille**
			Agressif	**GdPoids** Bull-Mast Dogue-All Mastiff Saint-Ber Terre-Neu	**Utile**		**Non-Affe**

TAB. 6.8 - *Carte de Kohonen simultanée; apprentissage avec la matrice K^c*

PtTaille **PtPoids** **PtVitesse** Basset Bull-Dog Chihuahua Pekinois Teckel	**Affect** **Compagnie** Caniche Cocker Fox-Terri	**MyTaille** **MyIntell** Boxer Dalmatien	**MyVitesse** **Non-Agre** Epagneu-B Labrador	**MyPoids** Colley Epagneu-F	**Chasse** Fox-Hound GrBlGasco Levrier Pointer Setter	**GrTaille** **GrVitesse** **PtIntell** **GrIntell** **Agressif** Beauceron Berge-All Doberman	**GdPoids** **Non-Affe** **Utile** Bull-Mast Dogue-All Mastiff Saint-Ber Terre-Neu

TAB. 6.9 - *Ficelle de Kohonen simultanée; apprentissage avec la matrice K^c*

6.4.2 L'algorithme KACM II : Apprentissage par la table de Burt

La seconde méthode consiste à utiliser uniquement les lignes de la matrice B^c pour faire l'apprentissage du réseau de Kohonen. C'est en fait la méthode KACM décrite dans 6.3 qui est utilisée ici, mais à la fin de l'apprentissage on classe les individus représentés par les vecteurs lignes de la matrice K^c, dans la carte de Kohonen. On regarde sur l'exemple des caractéristiques canines le résultat obtenu.

Les caractéristiques Canines

Sur la grille (8×8), on a quasiment les mêmes regroupements que ce qui précède mais de plus les classes plus homogènes et mieux séparées. Les classes sont mieux regroupées que celles de la carte 6.10. C'est dû, sans doute, au fait que l'inertie cumulée expliquée par les axes principaux de la matrice de Burt est plus grande que celle expliquée par les axes principaux du tableau disjonctif complet.

PtTaille **PtPoids** Chihuahua Pekinois Teckel	Bull-Dog		Basset	**PtIntell**			**GdPoids** Mastiff Saint-Ber Terre-Neu
Caniche Fox-Terri			**PtVitesse**			Dogue-All	Bull-Mast
Compagnie							**Utile**
Affectue	Cocker	**MyIntell**			**Agressif**		
MyTaille Boxer Dal-matien Labrador	**MyVitesse** Epagneu-B			**GrIntell** Beauceron Berge-All			**Non-Affe**
				Pointer	Doberman		
	Non-Agre Colley		**Chasse** Epagneu-F GrBlGasco Levrier Setter	**MyPoids**	**GrVitesse** Fox-Hound		**GrTaille**

TAB. 6.10 - *Carte de Kohonen simultanée; l'apprentissage est fait avec les lignes de la matrice corrigée* B^c

On le voit plus netement ici dans la ficelle (6.11) où l'on a la classe des chiens de compagnie qui se regroupent tous dans la même unité avec une unité quasi vide (Affectueux) qui la sépare des unités de droite. La même remarque vaut pour l'unité contenant les chiens de «grande taille», située à l'extrémité droite de la ficelle.

En comparaison, les figures 6.7 et 6.8 montrent les mêmes relations que celles obtenues sur les projections (1,2) et (1,3) dans une ACM classique.

Cette méthode est très intéressante et très économique du point de vue du temps de calcul. En effet, si on prend le cas de grands fichiers de données comme il en existe dans

PtTaille **PtPoids** **PtVitesse** **Com-pagnie** Basset Bull-Dog Caniche Chihuahua Cocker Fox-Terri Pekinois Teckel	**Affectue**	**MyTaille** **MyVi-tesse** **MyIntell** **Non-Agre** Boxer Dalmatien	Colley Epagneu-B Epagneu-F Labrador	**MyPoids** **Chasse** GrBlGasco Setter	**GrVitesse** **PtIntell** **GrIntell** Beauceron Berge-All Fox-Hound Levrier Pointer	**Agressif** Doberman	**GrTaille** **GdPoids** **Non-Affe** **Utile** Bull-Mast Dogue-All Mastiff Saint-Ber Terre-Neu

Tab. 6.11 - *Ficelle de Kohonen simultanée; l'apprentissage est fait avec les lignes de la matrice corrigée* B^c

les grandes enquêtes ou dans les fichiers de compagnies d'assurance. Il arrive que dans ce type de fichier, on ait des tableaux disjonctifs complets avec une centaine de variables mais des centaines de milliers (voir plus) d'individus. Une classification hiérarchique pour classer les individus en groupes peut prendre plusieurs heures. Par cette méthode, il suffit de calculer la matrice de Burt et d'entrainer un réseau de Kohonen au moyen de ses lignes. Comme généralement le nombre de modalités ne dépasse pas quelques centaines, l'apprentissage est très rapide. Il suffit alors de classer les individus dans la carte en recherchant les unités gagnantes en utilisant les lignes de la matrice K^c.

On peut également utiliser cette méthode pour les tableaux disjonctifs complets ayant des valeurs manquantes. On calcule la matrice de Burt restreinte aux observations présentes, on fait l'apprentissage d'un réseau de Kohonen avec la matrice de Burt corrigée correspondante, puis on classe les individus par rapport aux lignes de la matrice K^c. Il est évident que cette matrice est une matrice approchée. Il est possible de regrouper ensuite les unités en faisant une classification hiérarchique (ou autre) sur les vecteurs codes des unités pour construire des super classes.

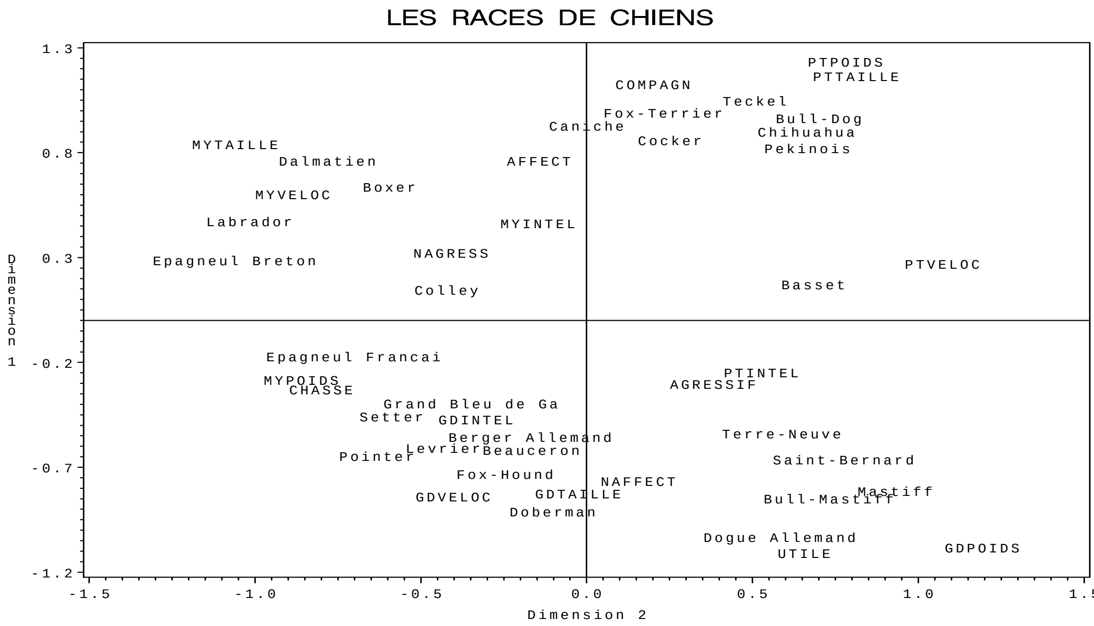


FIG. 6.7 - *Projection sur les axes 1 et 2*

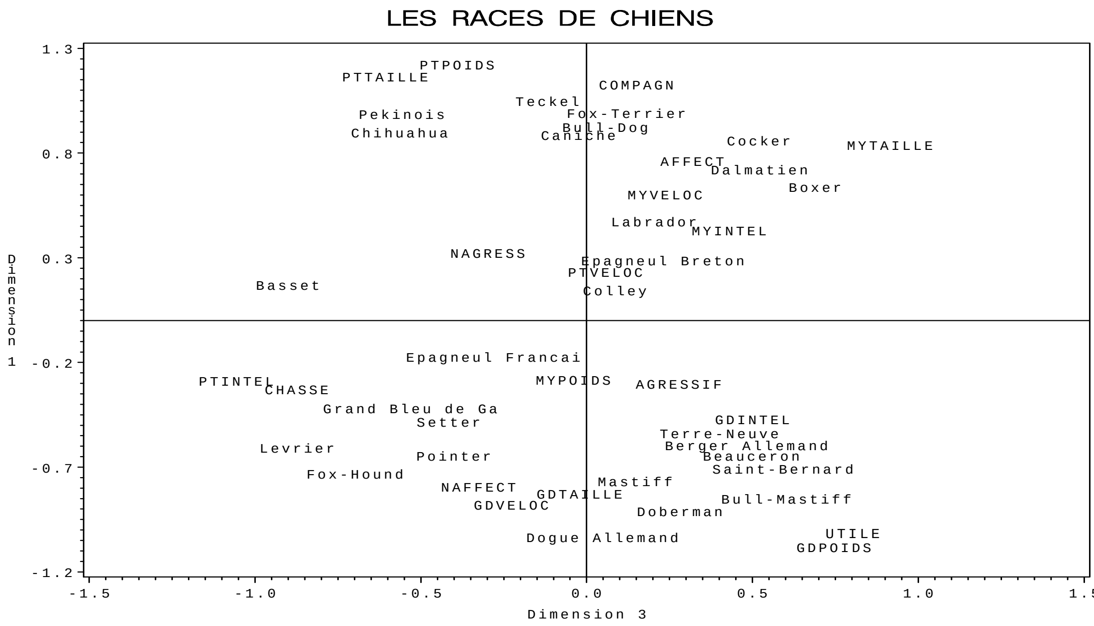


FIG. 6.8 - *Projection sur les axes 1 et 3*

6.5 Conclusion

Les résultats de cet algorithme sont très satisfaisants et prometteurs : en effet, sur les exemples pour lesquels on a appliqué l'algorithme **KACM**, on a obtenu rapidement une très bonne représentation des relations entre les variables. Cet algorithme présente l'avantage de comparer les profils en n'utilisant que la distance entre ces profils.

C'est une méthode de classification naturelle qui n'utilise pas d'approximation linéaire qui peut rendre parfois difficile l'interprétation des résultats. Les cartes qu'on obtient sont moins précises que les projections classiques de l'ACM, mais elles résument très bien les différentes relations (attractions, répulsions) entre modalités. Contrairement à ce qu'il pourrait paraître, la méthode est robuste puisqu'elle reproduit bien les classes stables.

Même si la représentation simultanée des différentes modalités des variables et des individus ne jouit pas d'une justification rigoureuse, néanmoins elle donne de bons résultats. Les méthodes **KACM I** et **KACM II** présentent l'avantage de produire d'une part une carte de Kohonen comparable à une projection d'une ACM classique et d'autre part une classification des individus par rapport aux modalités. De plus la méthode **KACM II** est particulièrement très économique du point de vue du temps de calcul, ce qui est un avantage considérable par rapport à d'autres méthodes classiques telles qu'une classification hiérarchique.

On peut également regrouper les unités du réseau en super-classes, en regroupant les vecteurs codes comme pour KACP. De même, il est possible d'utiliser la méthode **KACM II** sur des tableaux disjonctifs complets ayant des données manquantes.

Chapitre 7

Conclusion

Durant l'étude de l'algorithme de fusion, au chapitre 2, on a pu constater que la question du choix du nombre de classes à fixer dans une classification est un problème complexe qui n'a pas forcément de solution générale. En effet, le nombre de classes que l'on cherche à identifier dépend souvent de l'application pour laquelle on fait une classification. Dans le cas où l'on n'a pas d'idée a priori sur le nombre de classes, l'algorithme de Wong Y. F. paru dans (Wong, 1993) donne un critère assez simple pour le choix du nombre pertinent de classes. L'étude que nous avons menée nous a permis de la considérer comme une application concrète du principe d'estimation par maximum d'entropie. Cette étude peut servir de modèle pour l'élaboration de nouveaux algorithmes basés sur ce principe. En fait, on peut envisager d'utiliser d'autres contraintes que celle qui a été utilisée dans cet algorithme de fusion. On peut utiliser par exemple une contrainte prenant en compte une distance inter-classes maximale. L'algorithme de fusion reste malgré tout un algorithme lent par rapport à l'algorithme k-means mais probablement plus rapide, dans le cas de grande base de données, qu'une classification hiérarchique. Cependant un bon choix des deux paramètres qui gèrent cet algorithme améliore ses performances.

Nous avons ensuite rappelé le principe et quelques propriétés de convergence de l'algorithme d'auto-organisation de Kohonen. L'utilisation de cet algorithme pour remplacer une ACP et une classification simultanées a été développée dans trois exemples avec des données socio-économiques réelles. Nous avons proposé des techniques d'interprétation des cartes et de regroupement de classes. Il est clair que l'aspect non linéaire de l'algorithme représente un avantage qui peut être déterminant dans le choix de cette méthode (Carte Auto-Organisante) devant d'autres techniques linéaires telles que l'ACP. Mais on a aussi constaté une propriété encore plus intéressante de l'algorithme et qui est sa robustesse en présence de valeurs manquantes. Après avoir décrit le principe très simple de la méthode, nous avons mené une étude empirique afin d'évaluer le degré de cette robustesse. On a pu alors constater que l'algorithme de Kohonen adapté pour être utilisé avec des données manquantes, réalise de bons classements jusqu'à une proportion de quarante % de valeurs manquantes. Nous avons ensuite proposé une méthode d'estimation des valeurs manquantes par les valeurs des vecteurs poids, qui est beaucoup moins pénalisante qu'une estimation grossière par la moyenne globale sur les observations présentes.

Ce qui a été réalisé précédemment concerne des variables continues. Pour les variables qualitatives nous avons proposé la méthode **KACM**, basée sur l'algorithme d'auto-organisation de Kohonen, pour classer les modalités de Q variables qualitatives ou catégorielles, avec $Q \geq 2$. La carte obtenue résume les projections que génère une ACM classique. Les associations de modalités qui s'attirent fortement sont souvent regroupées dans de mêmes unités du réseau. La méthode **KACM** donne, de deux façons différentes, la possibilité de représenter les individus et les modalités des variables sur une même carte de Kohonen. On a alors une classification des individus et une carte comparable aux projections de type ACM.

L'ensemble de ces méthodes est actuellement utilisé dans des études économiques et financières de manière très satisfaisante. Elles apparaissent clairement comme des compléments très utiles des méthodes classiques.

Cependant les limitations de ces méthodes viennent du fait qu'on ne dispose pas actuellement de suffisamment de connaissance sur les propriétés de convergence de l'algorithme de Kohonen et le degré d'organisation d'une carte de Kohonen. En effet, pour le moment il n'existe que des critères approchés et visuels pour évaluer le degré d'organisation d'un réseau de Kohonen (Zrehen and Blayo, 1992), (Demartines, 1992), (Fort and Pagès, 1995) et (Villmann et al., 1997)). Bien que que l'intérêt de ces méthodes soit justement prouvé par leur utilisation pratique, un de nos objectifs est d'établir une méthodologie plus claire sur le choix de la taille du réseau et de sa géométrie ainsi qu'un indice quantifié des performances du réseau.

Bibliography

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks*, 2:53–58.

Benzecri, J. P. (1973). *L'analyse des données, T.2, L'analyse des correspondances*. Dunod, Paris.

Blayo, F. and Demartines, P. (1991). Data analysis: How to compare Kohonen neural networks to other technics? In Prieto, editor, *Proc of IWANN 91*, Lectures Notes in Computer Science, pages 469–476. Springer-Verlag.

Bock, H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. In Arabie, P., Hubert, L., and Soete, G. D., editors, *Clustering and classification*, pages 377–453. World Science, River Edge/NJ.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multi-layer perceptrons and singular value decomposition. *Biol. Cyb.*, 59:291–294.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture. *Journal of Classification*, 13(2):195–212.

Cottrell, M. and Fort, J. (1987). Etude d'un algorithme d'auto-organisation. *Ann. Inst. Henri Poincarré*, 23(1):1–20.

Cottrell, M., Fort, J., and Pages, G. (1995). Two or three things that we know about Kohonen algorithm. In Verleysen, M., editor, *Proc of ESANN'94*. D facto Bruxelles.

Cottrell, M., Girard, B., and Rousset, P. (1997). Forecasting of curves using a Kohonen Classification. *Journal of Forecasting*. to appear in.

Cottrell, M., Letremy, P., and Roy, E. (1993). Analyzing a contingency table with Kohonen maps: a Factorial Correspondence Analysis. In Cabestany, J., Mira, J., and Prieto, A., editors, *Proc of IWANN 93*, pages 305–311. Springer-Verlag.

Csiszar, I. (1970). I-Divergence Geometry Of Probability Distributions And Minimization Problems. *Ann. Prob.*, 3(1):146–159.

Dacunha-Castelle, D. and Gamboa, F. (1990). Maximum d'entropie et problème des moments. *Ann. Inst. Henri Poincaré*, 26(4):567–596.

Demartines, P. (1992). Organization measures and representation of Kohonen maps. In J.Hérault, editor, *First IFIP Working Group 10.6 Workshop*.

Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*. Wiley-interscience.

Erwin, E., Obermayer, K., and Shulten, K. (1992a). Self-organizing maps : ordering, convergence properties and energy functions. *Biol. Cyb.*, 67:47–55.

Erwin, E., Obermayer, K., and Shulten, K. (1992b). Self-organizing maps : stationary states, metastability and convergence rate. *Biol. Cyb.*, 67:35–45.

Földiák, P. (1989). Adaptative network for optimal linear feature extraction. In *Proceedings of the International Joint Conference on Neural Networks*, volume I, pages 401–405. San Diego : SOS Printing.

Fort, J. and Pagès, G. (1995). About the Kohonen algorithm : strong or weak self-organisatioon. *Neural Networks*, 9(5):773–785.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press Inc, 1250 Sixth Avenue, San Diego, CA 92101, 2nd edition edition.

Gallinari, P., Thiria, S., and Fogelman-Soulié, F. (1988). Multilayer Perceptrons and Data Analysis. In *Proceedings of the the 2nd ICNN*, volume I, pages 391–399, San Diego.

Gaubert, P., Ibbou, S., and Tutin, C. (1996). Segmented of Real Estate Markets and Price Mechanisms; The Case of Paris. *International Journal of Urban and Regional Research*, 20(2):270–298.

Gordon, A. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A*, 150:119–137.

Gordon, A. (1994a). Computational statistics : Papers collected on the occasion of the 25th conference on statistical computing at schloss reisenberg, contributions to statistics. Physica-Verlag.

Gordon, A. (1994b). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, 18(5):561–581.

Gordon, A. (1996). A survey of constrained classification. *Computational Statistics and Data Analysis*, 21(1):17–29.

Hartigan, J. (1975). *Clustering Algorithms*. Wiley, New York.

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation. Lecture*, volume 1. Addison-Wesley, santa fe institute edition.

Ibbou, S. and Cottrell, M. (1995a). Analyse des Correspondances Multiples sur tableau de burt au moyen de l'algorithme de Kohonen. In *XXVII journées de statistique*, Jouy-en-Josas (France).

Ibbou, S. and Cottrell, M. (1995b). Multiple Correspondence Analysis of a crosstabulations matrix using the Kohonen algorithm. In Verleysen, M., editor, *Proc of ESANN'95*, pages 27–32. D facto Bruxelles.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data.* Prentice-Hall, Englewood Cliffs.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev*, 106:620–630.

Kohonen, T. (1974). An adaptive associative memory principle. *IEEE Transactions on Computers*, C-23:444–445.

Kohonen, T. (1977). *Associative memory- A System Theoretical Approach.* Springer-Verlag, New York.

Kohonen, T. (1980). *Content-Addressable Memories.* Springer-Verlag, Heidelberg.

Kohonen, T. (1981). Automatic formation of topological maps in a self-organizing system. In Oja, E. and Simula, O., editors, *Proc. of the 2nd Scandinavian Conference of Image Analysis*, pages 214–220. Espoo: Suomen Hahmonttunnistustutkimujsen Seuro.

Kohonen, T. (1982). A simple paradigm for the self-organized formation of structured feature maps. In Arbib, M. and Amari, S., editors, *Lecture Notes in Biomathematics 45: Competition and Cooperation in Neural Nets*, pages 248–266. Berlin: Springer-Verlag.

Kohonen, T. (1984). *Self-organization and associative memory.* Springer, New York Berlin Heidelberg. 3^{rd} edition 1989.

Kohonen, T. (1986). Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, Departement of Technical Physics.

Kohonen, T. (1988). Learning vector quantization. *Neural Networks Supplement: INNS*, 1:303–310.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

Lebart, L., Morineau, A., and Piro, M. (1995). *Statistique exploratoire multidimensionnelle.* Dunod, Paris.

Letremy, P. (1997). Notice d'installation et d'utilisation de programmes basés sur l'algorithme de Kohonen et dédiés à l'analyse des données. Technical Report Prépublication du SAMOS n 82, Université Paris I Panthéon-Sorbonne, Laboratoire SAMOS.

Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for Vector Quantizer Design. *IEEE Transactions on communications*, Com-28(1):84–95.

Lloyd, S. P. (1957). Least Squares Quantization in PCM's. Technical report, Bell Telephone Laboratories Paper. Murray Hill.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the Fifth Berkeley Symposium on Math., Stat. and Prob.*, volume 1, pages 281–296.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematics and Biology*, 15:267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *Int. J. Neural Systems*, 1:61–68.

Ritter, H., Martinetz, T., and Shulten, K. (1992). *Neural computation and Self Organizing Maps, an Introduction.* Addison-Wesley, Reading.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer feedforward neural network. *Neural Networks*, 2:459–473.

Saporta, G. (1990). *Probabilités Analyses des données et statistique.* Editions Technip, Paris.

Villmann, T., Der, R., and Martinetz, T. (1997). Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurment. *IEEE tr. on Neural Networks*, 8(2):256–266.

Wong, Y. F. (1993). Clustering data by melting. *Neural computation*, 5(1):89–104.

Zrehen, S. and Blayo, F. (1992). A geometric organization measure for Kohonen's map. In *Proc of Neuro-Nîmes*, pages 603–610.

Annexes

Annexe A

Rappels d'algèbre linéaire

Dans ce paragraphe $\mathcal{E}$ et $\mathcal{F}$ désignent deux espaces vectoriels sur un corps K de dimensions finies respectives p et q. On suppose que les deux espaces sont respectivement munis des bases $\mathcal{B} = (e_1, ..., e_p)$ et $\mathcal{C} = (c_1, ..., c_q)$. Les vecteurs sont supposés être des vecteurs colonnes et si on note x les composantes d'un vecteur dans une base $\mathcal{B}$ alors x^t est le vecteur ligne transposé de x

Changement de base : Si $\mathcal{B}'$ désigne une autre base de $\mathcal{E}$, alors la matrice de passage P de la base $\mathcal{B}$ à $\mathcal{B}'$ est la matrice dont les colonnes sont les coordonnées des vecteurs de $\mathcal{B}'$ dans $\mathcal{B}$. Les relations de transformations de coordonnées sont

$$x' = Px \quad \text{et} \quad x = P^{-1}x'$$

où x et x' sont les coordonnées d'un vecteur de $\mathcal{E}$ respectivement dans la base $\mathcal{B}$ et $\mathcal{B}'$

Application linéaire : Si f est une application linéaire de $\mathcal{E}$ vers $\mathcal{F}$ alors sa matrice A par rapport aux bases $\mathcal{B}$ et $\mathcal{C}$ est la matrice dont les colonnes sont les images des vecteurs de la base $\mathcal{B}$ dans la base $\mathcal{C}$.
Si P est la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$ dans $\mathcal{E}$ et Q la matrice de passage de $\mathcal{C}$ à $\mathcal{C}'$ dans $\mathcal{F}$ et A' désigne la matrice de f par rapport à $\mathcal{B}'$ et $\mathcal{C}'$ alors

$$A' = P^{-1}AQ$$

Espace dual : Une forme linéaire h est une application linéaire de $\mathcal{E}$ vers K. Exemple : à tout vecteur de base e_i de $\mathcal{B}$ correspond une forme linéaire e_i^* qui à tout vecteur de $\mathcal{E}$ lui associe sa i-ème composante dans $\mathcal{B}$.

$$e_i^* \left| \begin{array}{ccc} \mathcal{E} & \longrightarrow & K \\ x & \longrightarrow & x_i \end{array} \right.$$

L'espace dual $\mathcal{E}^*$ de $\mathcal{E}$ est l'ensemble des formes linéaires sur $\mathcal{E}$. C'est un espace vectoriel de dimension p dont une base est :

$$\mathcal{B}^* = (e_1^*, ..., e_i^*, ..., e_p^*)$$

La base $\mathcal{B}^*$ est dite base duale de $\mathcal{B}$.

Application bilinéaire : Une application f de $\mathcal{E} \times \mathcal{F}$ vers $\mathcal{G}$

$$f \left| \begin{array}{ccc} \mathcal{E} \times \mathcal{F} & \longrightarrow & \mathcal{G} \\ (x, y) & \longrightarrow & f(x, y) \end{array} \right.$$

est dite bilinéaire si f est linéaire par rapport aux deux variables. Lorsque $\mathcal{G}$ est le corps K, f est dite *forme bilinéaire*. On lui associe la matrice $M = (a_{ij}) \in \mathcal{M}_{p\times q}$ telle que

$$\forall i, j \;\; 1 \leq i \leq p, \; 1 \leq j \leq q \;\; a_{ij} = f(e_i, c_j)$$

Alors

$$\forall (x, y) \in \mathcal{E} \times \mathcal{F} \qquad f(x, y) = x^t M y$$

On note $\mathcal{L}(\mathcal{E}, \mathcal{F}; \mathcal{K})$ l'ensemble des formes bilinéaires de $\mathcal{E} \times \mathcal{F}$ vers K. C'est un espace vectoriel isomorphe à $\mathcal{M}_{p\times q}$. Sa dimension est donc np. Si $\mathcal{F} = \mathcal{E}$, $\mathcal{L}(\mathcal{E}, \mathcal{F}; \mathcal{K})$ est noté simplement $\mathcal{L}_2(\mathcal{E}; \mathcal{K})$.

On suppose que $K = \mathbb{R}$, $\mathcal{L}_2(\mathcal{E}; \mathbb{R})$ est l'ensemble des formes bilinéaires réelles. Soit $f \in \mathcal{L}_2(\mathcal{E}; \mathbb{R})$ alors f est dite :

- symétrique si $\forall x, y \;\; f(x, y) = f(y, x)$
- positive si $\forall x \;\; f(x, x) \geq 0$
- définie si $f(x, x) = 0 \Longrightarrow x = 0$

A toute forme bilinéaire réelle f symétrique définie positive correspond une application q de $\mathcal{E}$ dans $\mathbb{R}^+$ appelée forme quadratique définie par :

$$q \left| \begin{array}{ccc} \mathcal{E} & \longrightarrow & \mathbb{R}^+ \\ x & \longrightarrow & q(x) = f(x, x) \end{array} \right.$$

Si M est la matrice définissant la forme bilinéaire f, alors M est symétrique ($M^t = M$) et

$$f(x, y) = x^t M y \qquad \text{et} \qquad q(x) = x^t M x$$

Cette forme bilinéaire définit un produit scalaire dans $\mathcal{E}$

- Le produit scalaire $< x, y >$ est : $< x, y >_M = x^t M y$
- La norme d'un vecteur x est : $\|x\|_M^2 = x^t M x$
- Un vecteur x de $\mathcal{E}$ est dit M-normé si : $\|x\|_M^2 = x^t M x = 1$
- Deux vecteurs x et y sont dits M-orthogonaux si : $< x, y >_M = x^t M y = 0$

Métrique (ou distance): La norme induite par la matrice M définit une distance dans l'espace vectoriel $\mathcal{E}$

$$d(x, y) = \|x - y\|$$

L'espace $\mathcal{E}$ est un espace euclidien.

PROPRIÉTÉ A.0.1

- Une application linéaire de $\mathcal{E}$ dans $\mathcal{E}$ définie par une matrice A de taille p est dite M-symétrique si $MA = A^tM$, c'est-à-dire que MA est symétrique

- Une matrice A de rang $q < p$ et M-symétrique a q valeurs propres strictement positives (éventuellement égales), une valeur propre 0 d'ordre de multiplicité $p - q$ et une base de vecteurs propres M-orthogonale.

Si M est la matrice identité de taille p, $(M = I_p)$, on retrouve le produit scalaire usuel

$$< x, y >= x^t y = \sum_{i=1}^{i=p} x_i y_i$$

Remarque1 (Décomposition de Choleski) :

Toute matrice réelle M symétrique définie positive se décompose en un produit $M = T^tT$ où T est une matrice inversible.

Ainsi

$$< x, y >_M = x^t T^t T y = (Tx)^t T y$$

Si on fait le changement de coordonnées $x' = Tx$, alors on retrouve le produit scalaire usuel sur les nouvelles coordonnées.

Annexe B

Liste des noms de communes

ACHE	ACHERES	MELU	MELUN
ACRU	ARCUEIL	MENN	MENNECY
ALFO	ALFORTVILLE	MEUD	MEUDON
ANDR	ANDRESY	MFER	MONTFERMEIL
ANTO	ANTONY	MIMO	MITRY-MORY
ARGE	ARGENTEUIL	MMAG	MONTMAGNY
ARNO	ARNOUVILLE-LES-GONESSE	MMOR	MONTMORENCY
ARPA	ARPAJON	MOIS	MOISSY-CRAMAYEL
ASNI	ASNIERES-SUR-SEINE	MONB	MONTIGNY-LE-BRETONNEUX
ATHI	ATHIS-MONS	MONG	MONTGERON
AUBG	AUBERGENVILLE	MONT	MONTEREAU-FAUT-YONNE
AUBV	AUBERVILLIERS	MORA	MORANGIS
AUFF	AUFFARGIS	MORS	MORSANG-SUR-ORGE
AULN	AULNAY-SOUS-BOIS	MRGE	MONTROUGE
AUVE	AUVERS-SUR-OISE	MSON	MONTESSON
AVON	AVON	MTRE	MONTREUIL
BAGT	BAGNOLET	MURO	MUREAUX
BEAU	BEAUCHAMP	NANT	NANTERRE
BEZO	BEZONS	NEMO	NEMOURS
BGLR	BOURG-LA-REINE	NEUM	NEUILLY-SUR-MARNE
BGNX	BAGNEUX	NEUP	NEUILLY-PLAISANCE
BLAN	BLANC-MESNIL	NEUS	NEUILLY-SUR-SEINE
BOBI	BOBIGNY	NOGM	NOGENT-SUR-MARNE
BOIA	BOIS-D'ARCY	NOIG	NOISY-LE-GRAND
BOIC	BOIS-COLOMBES	NOIL	NOISIEL
BOIL	BOISSY-SAINT-LEGER	NOIS	NOISY-LE-SEC
BOND	BONDY	ORLY	ORLY
BONE	BONNEUIL-SUR-MARNE	ORME	ORMESSON-SUR-MARNE
BOUL	BOULOGNE-BILLANCOURT	ORSA	ORSAY
BOUR	BOURGET	OSNY	OSNY
BRET	BRETIGNY-SUR-ORGE	OUEN	SAINT-OUEN-L'AUMONE
BRIC	SAINT-BRICE-SOUS-FORET	OZOI	OZOIR-LA-FERRIERE
BRIE	BRIE-COMTE-ROBERT	PA10	PARIS-10
BRUN	BRUNOY	PA11	PARIS-11
BRYM	BRY-SUR-MARNE	PA12	PARIS-12
BURE	BURES-SUR-YVETTE	PA13	PARIS-13
CACH	CACHAN	PA14	PARIS-14
CARP	CARRIERES-SOUS-POISSY	PA15	PARIS-15
CARS	CARRIERES-SUR-SEINE	PA16	PARIS-16
CBVE	COURBEVOIE	PA17	PARIS-17
CELL	CELLE-SAINT-CLOUD	PA18	PARIS-18
CERG	CERGY	PA19	PARIS-19
CHAN	CHANTELOUP-LES-VIGNES	PA20	PARIS-20
CHAT	CHATOU	PALA	PALAISEAU
CHAV	CHAVILLE	PANT	PANTIN
CHEL	CHELLES	PAR1	PARIS 1

CHEN	CHENNEVIERES-SUR-MARNE	PAR2	PARIS-2
CHES	CHESNAY	PAR3	PARIS-3
CHEV	CHEVILLY-LARUE	PAR4	PARIS-4
CHIL	CHILLY-MAZARIN	PAR5	PARIS-5
CHMA	CHATENAY-MALABRY	PAR6	PARIS-6
CHOI	CHOISY-LE-ROI	PAR7	PARIS-7
CHON	CHATILLON	PAR8	PARIS-8
CHPY	CHAMPIGNY-SUR-MARNE	PAR9	PARIS-9
CHSM	CHAMPS-SUR-MARNE	PARI	PARIS
CHTN	CHARENTON-LE-PONT	PAVB	PAVILLONS-SOUS-BOIS
CLAM	CLAMART	PECQ	PECQ
CLAY	CLAYES-SOUS-BOIS	PERS	PERSAN
CLIB	CLICHY-SOUS-BOIS	PERX	PERREUX-SUR-MARNE
CLIC	CLICHY	PERY	PERRAY-EN-YVELINES
COIG	COIGNIERES	PIER	PIERREFITTE-SUR-SEINE
COLO	COLOMBES	PLAI	PLAISIR
COMB	COMBS-LA-VILLE	PLES	PLESSIS-ROBINSON
CONF	CONFLANS-SAINTE-HONORINE	PLET	PLESSIS-TREVISE
CORB	CORBEIL-ESSONNES	POIS	POISSY
CORM	CORMEILLES-EN-PARISIS	PONT	PONTAULT-COMBAULT
COUL	COULOMMIERS	PORC	PORCHEVILLE
COUR	COURCOURONNES	PREG	PRE-SAINT-GERVAIS
CRET	CRETEIL	PROV	PROVINS
CROS	CROSNE	PTOI	PONTOISE
DAMG	DAMMARTIN-EN-GOELE	PUTO	PUTEAUX
DAML	DAMMARIE-LES-LYS	RAIN	RAINCY
DEUI	DEUIL-LA-BARRE	RAMB	RAMBOUILLET
DOMO	DOMONT	RISO	RIS-ORANGIS
DOUR	DOURDAN	ROIB	ROISSY-EN-BRIE
DRAN	DRANCY	ROIF	ROISSY-EN-FRANCE
DRAV	DRAVEIL	ROMA	ROMAINVILLE
EAUB	EAUBONNE	RONY	ROSNY-SOUS-BOIS
ELAN	ELANCOURT	RUEI	RUEIL-MALMAISON
ENGH	ENGHIEN-LES-BAINS	RUNG	RUNGIS
EPIO	EPINAY-SUR-ORGE	SACL	SACLAY
EPIS	EPINAY-SOUS-SENART	SANN	SANNOIS
EPNS	EPINAY-SUR-SEINE	SARC	SARCELLES
ERAG	ERAGNY	SART	SARTROUVILLE
ERMO	ERMONT	SAVI	SAVIGNY-LE-TEMPLE
ESSA	ESSARTS-LE-ROI	SAVO	SAVIGNY-SUR-ORGE
ETAM	ETAMPES	SCAU	SCEAUX
EVRY	EVRY	SCLO	SAINT-CLOUD
FONB	FONTENAY-SOUS-BOIS	SCYR	SAINT-CYR-L'ECOLE
FONF	FONTENAY-LE-FLEURY	SDEN	SAINT-DENIS
FONR	FONTENAY-AUX-ROSES	SEVR	SEVRES
FONT	FONTAINEBLEAU	SGEN	SAINTE-GENEVIEVE-DES-BOIS
FRAN	FRANCONVILLE	SGER	SAINT-GERMAIN-EN-LAYE
FRES	FRESNES	SLEU	SAINT-LEU-LA-FORET
GAGN	GAGNY	SMAN	SAINT-MANDE
GARC	GARCHES	SMAU	SAINT-MAUR-DES-FOSSES
GARE	GARENNE-COLOMBES	SMIC	SAINT-MICHEL-SUR-ORGE
GARG	GARGES-LES-GONESSE	SMOR	SAINT-MAURICE
GARV	GARGENVILLE	SOIS	SOISY-SOUS-MONTMORENCY
GENN	GENNEVILLIERS	SOUE	SAINT-OUEN
GENT	GENTILLY	SQYa	SQY
GIFS	GIF-SUR-YVETTE	STAI	STAINS
GONE	GONESSE	STFA	SAINT-FARGEAU-PONTHIERRY
GOUS	GOUSSAINVILLE	STPI	SAINT-PIERRE-LES-NEMOURS
GRAT	SAINT-GRATIEN	SUCY	SUCY-EN-BRIE
GRIG	GRIGNY	SURE	SURESNES
GROS	GROSLAY	SVRA	SEVRAN
GUYA	GUYANCOURT	TAVE	TAVERNY

HAYR	HAY-LES-ROSES	THIA	THIAIS
HERT	HERBLAY	TORC	TORCY
HOUD	HOUDAN	TRAP	TRAPPES
HOUI	HOUILLES	TREF	TREMBLAY-EN-FRANCE
IGNY	IGNY	ULIS	ULIS
ISLE	ISLE-ADAM	VAIR	VAIRES-SUR-MARNE
ISSY	ISSY-LES-MOULINEAUX	VALE	VALENTON
ISTD	ILE-SAINT-DENIS	VANV	VANVES
IVRY	IVRY-SUR-SEINE	VAUC	VAUCRESSON
JOIN	JOINVILLE-LE-PONT	VAUR	VAUREAL
JOUY	JOUY-LE-MOUTIER	VAVR	VILLE-D'AVRAY
JUVI	JUVISY-SUR-ORGE	VBEL	VILLIERS-LE-BEL
KREM	KREMLIN-BICETRE	VELI	VELIZY-VILLACOUBLAY
LACO	COURNEUVE	VERB	VERRIERES-LE-BUISSON
LAGM	LAGNY-SUR-MARNE	VERI	VERRIERE
LESI	LESIGNY	VERN	VERNEUIL-SUR-SEINE
LEVA	LEVALLOIS-PERRET	VERS	VERSAILLES
LILA	LILAS	VESI	VESINET
LIMA	LIMAY	VGAR	VILLENEUVE-LA-GARENNE
LIME	LIMEIL-BREVANNES	VIGS	VIGNEUX-SUR-SEINE
LIVR	LIVRY-GARGAN	VILM	VILLIERS-SUR-MARNE
LOGN	LOGNES	VILP	VILLEPARISIS
LONG	LONGJUMEAU	VINC	VINCENNES
MAGN	MAGNY-EN-VEXIN	VIPE	VILLEPREUX
MAIL	MAISONS-LAFFITTE	VIRO	VIROFLAY
MALA	MALAKOFF	VIRY	VIRY-CHATILLON
MALF	MAISONS-ALFORT	VITR	VITRY-SUR-SEINE
MANJ	MANTES-LA-JOLIE	VJUI	VILLEJUIF
MANV	MANTES-LA-VILLE	VMOM	VILLEMOMBLE
MARL	MARLY-LE-ROI	VOIS	VOISINS-LE-BRETONNEUX
MARN	MARNES-LA-COQUETTE	VPIN	VILLEPINTE
MASS	MASSY	VROI	VILLENEUVE-LE-ROI
MAUR	MAUREPAS	VSTG	VILLENEUVE-SAINT-GEORGES
MCOR	MONTIGNY-LES-CORMEILLES	VTAN	VILLETANEUSE
MEAU	MEAUX	YERE	YERRES
MEES	MEE-SUR-SEINE		

Annexe C

Article : International Journal of Urban and Regional Research

Annexe D

Article : Cahiers Économiques de Bruxelles