La méthode de classification non-supervisée K-means axiales

Alain Lelu

A. Lelu – alain.lelu@univ-fcomte.fr
Professeur
Université de Franche Comté / LASELDI
LORIA / Equipe KIWI

Document de référence (26/03/2004 + corrections 2006, 2008)

La méthode de classification non-supervisée K-means axiales (KMA)

Plan

1. L'espace des K-means axiales

Cette méthode est basée sur le principe de classification par centres mobiles, plus connue sous le nom de K-means – cf. [Forgy 65] pour sa variante non adaptative et [MacQueen 67] pour sa

variante adaptative. La variante d'Alain Lelu, les K-means axiales, réalisant une analyse factorielle de chaque classe dégagée dans l' « espace distributionnel », il est donc dans un premier temps intéressant de se faire une idée de cet espace et de cette analyse factorielle.

1.1. Distance et cosinus distributionnels

Toute méthode d'analyse des données est caractérisée à la base par trois choix : 1) une transformation opérée sur les vecteurs-données bruts, 2) une métrique, ou pondération des dimensions dans lesquels ces vecteurs sont définis, 3) une pondération de ces vecteurs. Sur le nuage de points ainsi défini, de nombreuses techniques de synthèse d'information et réduction des dimensions peuvent être appliquées : classification ascendante ou descendante hiérarchique, classification à centres mobiles, décomposition aux valeurs singulières, incluant toutes les méthodes d'analyse factorielle qui en sont les variantes. Notre méthode des K-means axiales n'échappe pas à la règle ; voyons tout d'abord la transformation des données et la métrique choisies, qui définissent notre distance :

Une lignée ancienne de travaux [Matusita 1955] [Escofier 1978] [Domengès et Volle 1979] [Fichet et Gbegan 1985] s'est intéressée à ce que ces auteurs appellent *distance distributionnelle* (ou encore : distance de Hellinger) :

La distance distributionnelle est la distance euclidienne, classique (équipondération des dimensions), entre les 2 points t1 et t2, de coordonnées les vecteurs $\mathbf{z}_{t1}$ et $\mathbf{z}_{t2}$, situés sur l'hypersphère unité dans l'espace des I mots et représentant chacun une unité textuelle, définis par la transformation suivante sur les données :

$\mathbf{z}_{t1} : \{\sqrt{}(x_{it1}/ x_{.t1}) \}$; $\mathbf{z}_{t2} : \{\sqrt{}(x_{it2}/ x_{.t2}) \}$

où x_{it} désigne la fréquence du mot i dans le document t, et $x_{.t}$ le nombre total de mots du document t

La distance distributionnelle Dd(t1, t2) entre les textes t1 et t2 est donc :

$$Dd(t1, t2) = \| \mathbf{z}_{t1} - \mathbf{z}_{t2} \|$$

où $\| \mathbf{x} \|$ désigne la norme euclidienne du vecteur $\mathbf{x}$.

Cette distance est la longueur de la corde correspondant à l'angle ($\mathbf{z}_{t1}$, $\mathbf{z}_{t2}$) - égale au plus à 2 quand ces 2 vecteurs sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi cette normalisation insolite plutôt que la normalisation classique $\{x_{it}/ \|\mathbf{x}_t\| \}$?) , mais elle jouit de propriétés intéressantes :

- Contrairement à la distance du khi-deux utilisée en Analyse Factorielle des Correspondances (AFC), elle peut prendre en compte des vecteurs ayant des composantes négatives, propriété utile pour certains types de codage « symétriques » (comme *Oui, Non, Ne sait pas*) ou pour des tableaux de flux orientés – économiques, physiques, ...
- Elle est liée à la mesure du gain d'information de Renyi d'ordre ½ [Renyi 1966] apporté par une distribution $\mathbf{x}_q$ quand on connaît la distribution $\mathbf{x}_p$:
$$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = - 2 \log_2 (\cos(\mathbf{z}_p, \mathbf{z}_q)) = - 2 \log_2 (1 - Dd^2/2)$$
- Elle est rapide à calculer dans le cas des données textuelles, où les vecteurs $\mathbf{y}_t$ sont très « creux ».

- et surtout Escofier et Volle ont montré qu'elle satisfaisait à la même propriété d'équivalence distributionnelle que la distance du khi-deux utilisée en AFC : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, dans le cas où les descripteurs sont des mots et les unités décrites des textes, cette propriété assure la stabilité du système des distances entre textes au regard de l'éclatement ou du regroupement de mots de distributions proches.
- Nous avons confirmé empiriquement ces considérations théoriques à partir des données du défi Amaryllis 2003 de recherche d'information [Lelu 03].

1.2. Décomposition aux valeurs singulières dans l'espace distributionnel

Si l'on transforme le tableau (textes × mots) des données brutes comme suit :
. Coordonnées : - des vecteurs-lignes $\mathbf{x}_t : \{ x_{it} \} \rightarrow \mathbf{y}_t : \{ \sqrt{} x_{it} \}$
- des vecteurs-colonnes $\mathbf{x}_i : \{ x_{it} \} \rightarrow \mathbf{y}_i : \{ \sqrt{} x_{it} \}$
. Poids de ces vecteurs : unité
. Métrique : euclidienne standard
les cosinus entre vecteurs-ligne $\mathbf{y}_t$ (resp. entre vecteurs-colonnes $\mathbf{y}_i$) possèdent dans cet espace des propriétés intéressantes. Ils sont liés en effet à la notion de distance distributionnelle Dd par la relation :
$Dd(t_1,t_2)^2 = 2(1-\cos(t_1,t_2))$
(resp. $Dd(i_1,i_2)^2 = 2(1-\cos(i_1,i_2))$)

Si on calcule les k premières directions propres $\mathbf{U} = \{ \mathbf{u}^{(k)} \}$ (resp $\mathbf{W} = \{ \mathbf{w}^{(k)} \}$) du nuage des points-colonnes $\mathbf{y}_t$ (resp, des points-lignes $\mathbf{y}_i$) défini plus haut, au moyen de la décomposition aux valeurs singulières du tableau des racines carrées[1] des données, on démontre que les cosinus ci-après se déduisent des directions propres :
$\text{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} \sqrt{(\lambda^{(k)} / x_{.t})}$ (1)
$\text{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} \sqrt{(\lambda^{(k)} / x_{i.})}$ (2)

Ces cosinus peuvent être considérés comme les facteurs d'un cas particulier et simple d'analyse factorielle sphérique, pour reprendre la terminologie de M. Volle, dite centrée sur le « tableau nul ». Nous les nommerons désormais respectivement $F_t^{(k)}$ et $G_i^{(k)}$. Ils sont liés entre eux et avec les $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :
$F_t^{(k)} = \Sigma_i \, u_i^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$
$G_i^{(k)} = \Sigma_t \, w_t^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$
Ce qui permet de les calculer à partir de l'extraction des éléments propres de celui des deux nuages qui comporte le moins de points.

Le principe de notre algorithme K-means axiales [Lelu, 1994] est de réaliser la partition en classes des unités textuelles dans l'espace distributionnel, et d'extraire pour chaque classe les facteurs mots et documents définis ci-dessus. Chaque facteur est alors un indicateur de *centralité* (ou « typicité ») du texte (« document ») ou du mot dans sa classe.

[1] Pour des données négatives, racines des valeurs absolues, signées : cf. *infra.*

Contrairement à la distance du khi-deux [Benzecri et al., 1981] et au cosinus de Salton [Salton, 1983], les calculs de similarité dans cet espace ne font pas intervenir une métrique de type *inverse document frequency* (IDF). Ceci dit, il est à noter que le passage d'un ensemble de documents à l'ensemble des mots qui en sont les plus caractéristiques se fait en deux étapes :

1) Calcul du 1er vecteur propre du tableau des racines carrées des fréquences brutes,
2) Calcul des facteurs-mots via la formule (2), qui fait intervenir une correction de type IDF.

On peut donc dire que la pondération de type IDF intervient aussi dans l'espace distributionnel, mais de façon implicite, et non directe comme dans le cosinus de Salton ou la distance du khi-deux.

Ce type de représentation possède des propriétés intéressantes :

- L'inertie autour du premier axe issu de la décomposition, somme des carrés des projections pondérées des points sur l'axe, se compare à la somme totale du tableau et s'interprète en terme de « part de données décrite».
- Elle possède également la propriété de dualité entre l'analyse des lignes et des colonnes ce qui signifie que les représentations des lignes et des colonnes sont identiques, que l'analyse soit faite sur le nuage des vecteurs-lignes ou des vecteurs-colonnes [Lebart *et al.* 82].
- Le premier facteur de cette analyse permet la reconstitution optimale des données au sens précisé dans [Domengès & Volle 79] (approximation d'un tableau de données par un produit de deux vecteurs).
- Cette analyse se généralise à la prise en compte des nombres négatifs, ce qui élargit les possibilités de codage et les champs d'application, moyennant la transformation $\sqrt{\xi_{ti}} \leftarrow \left[signe\left(\xi_{ti}\right)\right]\sqrt{\left\|\xi_{ti}\right\|}$ et les conventions $\xi_{t.} = \sum_i \left\|\xi_{ti}\right\|$, $\xi_{.i} = \sum_t \left\|\xi_{ti}\right\|$ et $\xi_{..} = \sum_{t,i} \left\|\xi_{ti}\right\|$.

(cf. notations ci-dessous)

2. Présentation formelle de l'analyse

Soit $X = \left\{\xi_{ti}\right\}$ le tableau des données brutes de *T* lignes et *I* colonnes.

- $A = \left\{\sqrt{\xi_{ti}}\right\}$ le tableau transformé
- $a_t = \left[\sqrt{\xi_{t1}},...,\sqrt{\xi_{tI}}\right]$ vecteur ligne de A
- $a_i = \left[\sqrt{\xi_{1i}},...,\sqrt{\xi_{Ti}}\right]'$ vecteur colonne de A (u' désigne la transposée de *u*)
- $u = \left[u_1,...,u_I\right]'$ premier vecteur propre de $A'A$
- $w = \left[w_1,...,w_T\right]'$ premier vecteur propre de AA'

Il est possible d'interpréter ces deux dernières entités comme des axes pointant dans la direction d'inertie maximale de deux nuages de points dans l'espace des colonnes et des lignes respectivement en supposant les lignes et les colonnes équipondérées.

Ces vecteurs propres sont les mêmes que ceux des nuages de points transformés [Lelu 99]:

- $a_t^{(i)} = \left[\sqrt{\xi_{t1}/\xi_{t.}},...,\sqrt{\xi_{tI}/\xi_{t.}}\right]$ pondéré par $\xi_{t.}$ situés sur l'hypersphère dans l'espace $\Re^I$

- $a_i^{(t)} = \left[\sqrt{\xi_{1i}/\xi_{.i}},\ldots,\sqrt{\xi_{Ti}/\xi_{.i}}\right]'$ pondéré par $\xi_{.i}$ situés sur l'hypersphère dans l'espace $\Re^T$

Les facteurs φ_t et γ_i sont définis respectivement comme les projections des vecteurs normés $a_t^{(i)}$ et $a_i^{(t)}$ sur ces directions propres – ce sont donc des cosinus des angles qu'ils forment avec les directions propres. Trois couples de formules de transition permettent de calculer les composantes des vecteurs propres, ainsi que les facteurs, attachés aux lignes comme aux colonnes, à partir de n'importe quel élément :

$$u_i = \left(1/\sqrt{\lambda}\right)\sum w_t\sqrt{\xi_{ti}} \qquad w_t = \left(1/\sqrt{\lambda}\right)\sum u_i\sqrt{\xi_{ti}}$$

$$\varphi_t = \left(1/\sqrt{\lambda\xi_{t.}}\right)\sum_i \gamma_i\sqrt{\xi_{ti}\xi_{.i}} \qquad \gamma_i = \left(1/\lambda\xi_{.i}\right)\sum_t \varphi_t\sqrt{\xi_{ti}\xi_{t.}}$$

$$\varphi_t = w_t\sqrt{\lambda/\xi_{t.}} \qquad \gamma_i = u_i\sqrt{\lambda/\xi_{.i}}$$

La projection η_t de chaque ligne a_t sur la première direction propre est donnée par :

$$\eta_t = \langle a_t, u\rangle = w_t\sqrt{\lambda} = \varphi_t\sqrt{\xi_{t.}}$$

λ représente la première valeur propre, inertie du premier vecteur propre :

$$\lambda = \sum_t \eta_t^2$$

L'inertie totale du tableau des données n'est autre que la somme totale des données :

$$\psi^2 = \sum\sum\sqrt{\xi_{ti}}^2 = \sum\sum\xi_{ti} = \xi_{..}$$

Une approximation de premier ordre des données peut être obtenue à partir du premier facteur :

$$\xi_{ti} \cong w_t^2 u_i^2\lambda = \varphi_t^2\gamma_i^2\xi_{.i}\xi_{t.}/\lambda = u_i^2\eta_t^2$$

Contrairement à l'analyse factorielle des correspondances, où le premier facteur dit « facteur trivial » est négligé, l'intérêt de l'analyse factorielle sphérique porte sur ce facteur puisqu'il traduit la « densité angulaire » des vecteurs données. D'autre part, ce type d'analyse s'adapte particulièrement aux vecteurs données dont la norme importe peu, c'est-à-dire aux tableaux de correspondance (comptages d'occurrences) utiles à l'analyse de données documentaires et textuelles et ce du fait de la normalisation du tableau de données.

2.1 Algorithme des K-means axiales (KMA)

Cette méthode fait appel dans un premier temps à un algorithme de type agrégation autour de centres mobiles, la méthode des K-means, qui a pour principe de former des groupements d'individus en affectant les éléments à des centres provisoires de classes, puis en recentrant ces classes, et en affectant de façon itérative ces éléments. Dans un second temps, la méthode des K-means axiales réalise une analyse factorielle sphérique sur chaque classe dégagée par la classification de type centres

mobiles. De même que la méthode des K-means classique, la classification K-means axiales existe sous deux versions, une version adaptative qui effectue une mise à jour immédiate des centres des classes et une version non adaptative qui met à jour ces centres après passage de tous les individus. Dans le cas des K-means axiales, les centres des classes sont des demi axes représentatifs des individus.

Les étapes de l'algorithme KMA sous sa version adaptative (en une seule passe sur les données) sont les suivantes :

2.1.1 Initialisation

Choix au hasard de *K* demi axes $u(0)^{(k)} = \left[u_1(0)^{(k)}, ..., u_I(0)^{(k)} \right]$ normalisés : $\left\| u(0)^{(k)} \right\| = 1$

et de *K* scalaires : $\tau_0(k) = 0$; *k* est l'indice des classes (*k*=1,...,*K*)

2.1.2 Passage de *t*-1 à *t*

Pour chaque ligne $a_t = \left[\sqrt{\xi_{t1}}, \ldots, \sqrt{\xi_{tI}} \right]$ du tableau de données transformé :

. Calcul des *K* projections sur les demi axes $u^{(k)}$ soit $\eta_t^{(k)} = < u(t)^{(k)}, a_t >$

. a_t est intégré à la classe *k* pour laquelle sa projection $\eta_t^{(k)}$ est maximale

. $u^{(k)}$ est alors mis à jour par la loi d'Oja modifiée [Oja 82] [Lelu 99] : $u(t)^{(k)} = u(t-1)^{(k)} + \left(\eta_t^{(k)} / \tau_t(k) \right) \left(a_t - \eta_t^{(k)} u(t-1)^{(k)} \right)$ avec $\tau_t(k) = \tau_{t-1}(k) + \eta_t^{(k)^2}$ qui représente l'inertie de la classe *k*.

2.1.3 Test d'arrêt

L'algorithme se termine lorsque la totalité des vecteurs lignes a été parcourue et le calcul des valeurs factorielles (issues d'une analyse factorielle sphérique) donne :

- Facteurs colonne $\gamma_i^{(k)} = u_i^{(k)} \sqrt{\lambda^{(k)} / \xi_{.i}^{(k)}}$
- Facteurs ligne $\varphi_t = \eta_t^{(k)} / \sqrt{\xi_{t.}}$

Comme il a été vu dans la présentation de l'analyse factorielle sphérique minimale (§ 1.2), ces deux facteurs représentent des cosinus, à savoir la projection des points représentatifs des individus normalisés sur les directions propres issues de cette analyse. Ce qui constitue une façon de positionner les individus d'une classe les uns par rapport aux autres, par des valeurs de centralité dans la classe.

2.2 Erreur relative de reconstitution et indicateur de qualité

L'erreur relative de reconstitution des données s'écrit :

$$\varepsilon \cong \left(\xi_{..} - \sum_k \tau(k) \right) \Big/ \xi_{..}$$

La valeur de l'indicateur de qualité de l'analyse est par conséquent :

$$1 - \varepsilon \cong \sum_k \tau(k) \Big/ \xi_{..}$$

Cette valeur permet d'attribuer un critère quantitatif de qualité à la méthode K-means axiales et donc de déterminer, parmi plusieurs classifications KMA sous différentes graines d'initialisation et à K constant, laquelle regroupe les individus de la façon la plus homogène.

Avantage : Cet algorithme est rapide (O(TK)) et consomme très peu de mémoire (O(IK)). De plus, il est incrémental ce qui signifie que l'introduction d'un nouvel individu ne nécessite pas de reprendre toutes les étapes du processus.

Inconvénient : Le résultat de cet algorithme varie sensiblement 1) avec l'initialisation 2) avec l'ordre d'entrée des données.

2.3 La version non adaptative de l'algorithme

2.3.1 Initialisation

Choix au hasard de *K* demi axes $u(0)^{(k)} = \left[u_1(0)^{(k)}, \ldots, u_I(0)^{(k)} \right]$ normalisés : $\left\| u(0)^{(k)} \right\| = 1$

et de *K* scalaires : $\tau_0(k) = 0$; *k* est l'indice des classes (k=1,...,K)

2.3.2 Passage de *t*-1 à *t*

Pour chaque ligne $a_t = \left[\sqrt{\xi_{t1}}, \ldots, \sqrt{\xi_{tI}} \right]$ du tableau de données transformé

- Calcul des *K* projections sur les demi axes $u^{(k)}$ soit $\eta_t^{(k)} = < u(t)^{(k)}, a_t >$
- a_t est intégré à la classe *k* pour laquelle sa projection $\eta_t^{(k)}$ est maximale
- $\tau_t(k) = \tau_{t-1}(k) + \eta_t^{(k)^2}$
- le vecteur « accumulateur d'apprentissage » $u'^{(k)}$ représentatif de la classe est mis à jour $u(t)^{(k)} = u(t-1)^{(k)} + \left(\eta_t^{(k)} / \tau_t(k) \right) a_t$ et normalisé.

2.3.3 Test d'arrêt

La totalité des vecteurs lignes est parcourue. Les axes des classes sont remplacés par les $u^{(k)}$. Le calcul du critère $\tau = \sum_k \tau(k)$ définit le test d'arrêt. En effet, si ce critère augmente par rapport au passage précédent d'une quantité inférieure à un seuil paramétrable ou s'il n'y plus de changements d'affectation l'algorithme passe au calcul des valeurs factorielles :

- Facteurs colonne $\gamma_i^{(k)} = u_i^{(k)} \sqrt{\lambda^{(k)} / \xi_{.i}^{(k)}}$
- Facteurs ligne $\varphi_t = \eta_t^{(k)} / \sqrt{\xi_{t.}}$

L'interprétation de ces valeurs est la même que dans le version adaptative.

2.3.4 Erreur relative de reconstitution et indicateur de qualité

L'erreur relative de reconstitution des données s'écrit :

$$\varepsilon \cong \left(\xi_{..} - \sum_k \tau(k) \right) \Big/ \xi_{..}$$

La valeur de l'indicateur de qualité de l'analyse est par conséquent :

$$1 - \varepsilon \cong \sum_k \tau(k) \Big/ \xi_{..}$$

Avantage de l'algorithme : Il est rapide (O(TK)) et consomme peu de mémoire (O(IK)).

Inconvénient : Le résultat varie avec l'initialisation des *K* demi axes, car la fonction objectif $\tau = \sum_k \tau(k)$ converge vers des maxima locaux, en fonction de l'initialisation.

Une heuristique efficace pour accélérer le calcul et améliorer la qualité de l'analyse consiste à commencer le processus par une passe adaptative, qui amène l'inertie intra-classes proche d'un de ses maxima locaux et ensuite d'utiliser la version non adaptative de l'algorithme qui converge après plusieurs passes sur les données [Lelu 93]. Une autre voie consiste à utiliser l'heuristique d'initialisation « K-means++ » [Arthur & Vassilvitskii 06].

D'autre part, il est possible d'adapter cet algorithme à des *tableaux de données à valeurs négatives*, il suffit pour cela d'effectuer les changements suivants :

$$\sqrt{\xi_{it}} \leftarrow \left(signe\left(\xi_{it}\right) \right) \sqrt{\left\| \xi_{it} \right\|} \qquad \xi_{i.} \leftarrow \sum_t \left\| \xi_{it} \right\| \quad \xi_{.t} \leftarrow \sum_i \left\| \xi_{it} \right\| \quad \xi_{..} \leftarrow \sum\sum \left\| \xi_{it} \right\|$$

Les tableaux de données négatives donnent la possibilité de gérer des questionnaires d'enquêtes (Oui=1, Non=-1, Sans opinion=0), ainsi que des ensembles de données comportant des valeurs négatives dans le domaine de descriptions de flux économiques ou autres, ou d'utiliser des transformations de données donnant lieu à des valeurs négatives.

Il faut également préciser que le nombre de classes résultant des algorithmes précédents n'est pas nécessairement égal au nombre de classes choisies au départ. En effet, il se peut qu'aucun individu ne soit affecté à une classe : le nombre de demi axes choisis au départ correspond au nombre maximum de classes désiré.

3. Interprétation des valeurs factorielles

La direction du premier axe factoriel est proche du point $\left\{\sqrt{\xi_{i.}}\right\}$ [Domengès, Volle 79], qui est le transformé du centre de gravité $\left\{\xi_{i.}\right\}$ du nuage de points d'origine. Les valeurs du premier vecteur propre pour les colonnes sont donc ordonnées de façon voisine des proportions de chaque descripteur dans la sous-population formée par les individus d'une classe issue de la classification. Mais on peut également interpréter les facteurs-colonnes comme les centralités des nœuds du graphe (pondéré) dont les arêtes ont pour valeur les cosinus entre colonnes (resp. lignes). Ainsi cet indicateur fait ressortir les descripteurs dont les proportions au sein de la classe sont les plus constantes, quel que soit leur valeur absolue (par ex. des mots rares équirépartis dans la classe, sans concentration de leur fréquence sur un petit nombre de documents, aussi bien que des mots fréquents équirépartis auront une faible valeur).

L'information précédente (« intrinsèque ») permet de décrire la structure dominante dans la classe mais ne permet pas de la différencier par rapport à celle des autres classes ; c'est pourquoi, de façon à faire ressortir les descripteurs spécifiques d'une classe et rares dans l'ensemble, nous pondérons l'importance absolue d'un descripteur dans une classe par l'inverse de sa fréquence dans la base :

$\gamma_i^{(k)} = u_i^{(k)}\sqrt{\lambda^{(k)}/\xi_{i.}^{(k)}}$ devient $\rho_i^{(k)} = u_i^{(k)}\sqrt{\lambda^{(k)}/\xi_{i.}}$, en multipliant par $\sqrt{\xi_{i.}^{(k)}/\xi_{i.}}$.

Ce deuxième indicateur (« relatif ») réalise un compromis entre la répartition des descripteurs dans la classe tout en faisant ressortir les descripteurs fréquents dans une classe et rares dans l'ensemble : c'est le plus « parlant », celui que nous employons généralement pour prendre rapidement connaissance d'un corpus inconnu.

4. Considérations sur le recouvrement des thèmes

Il est possible de projeter n'importe quel individu sur l'axe représentatif d'une classe, ceci permet de faire apparaître des individus extérieurs à une classe mais qui peuvent se projeter très haut sur l'axe représentatif de cette classe, ce qui signifie que l'individu n'est pas loin de faire partie de cette classe. C'est pourquoi, en introduisant un seuil dit de *typicité*, il est possible d'obtenir des classes recouvrantes - cependant comme tout paramètre ce seuil influe sur les résultats de classification.

Il faut également remarquer que les axes représentatifs d'une classe ne sont pas nécessairement orthogonaux, à la différence des analyses factorielles classiques. Un avantage de cette considération angulaire est que l'ajout ou le retrait de vecteurs données ne modifie que localement la représentation obtenue. En effet, la théorie et la pratique des K-means axiales montrent que si les données correspondant à une certaine classe sont retranchées aux données analysées et que l'analyse est refaite pour *K*-1 classes, la représentation des classes restantes reste voisine [Lelu 93], à la variabilité résultant de l'initialisation près.

5. Dualité de l'analyse des individus et des descripteurs

Dans l'algorithme des KMA, qui utilise l'analyse factorielle sphérique minimale, la dualité entre analyse des lignes et des colonnes ne se situe qu'au niveau local, c'est-à-dire à l'intérieur de chacune des classes obtenues. Les formules de dualité vues plus haut sont valables en se restreignant à une classe *k*. Par contre, au niveau global, l'analyse sur les descripteurs et celle sur les individus n'ont aucune raison de se ressembler même si pratiquement, dans le domaine des données documentaires et textuelles qui sont très multidimensionnelles et à faible nombre de descripteurs présents par

description (données « pick-any » [Lelu 93]), on constate qu'à des classes d'individus correspondent des classes de descripteurs.

6. Stabilité de l'analyse

A nombre de classes *K* fixé, il est possible de varier les graines d'initialisation : à chacune d'elles correspond une valeur de critère de qualité de classification (la somme des inerties intra-classes). Cette procédure de recherche des « meilleures classes » auxquelles les utilisateurs puissent se fier est de nature différente, mais aussi fastidieuse que la recherche des « formes fortes » [Diday 79] communes à de nombreux passages sous différentes initialisations. De telles procédures sont inévitables dans les méthodes de classification à centres mobiles, qui réalisent des « quasi-optimisations » dépendantes des conditions initiales. C'est justement pour éviter de tels inconvénients qu'il est souhaitable de trouver des méthodes qui fournissent directement un optimum absolu et qui, par conséquent, pointent d'emblée sur les « formes fortes ». Nous poursuivons nos recherches sur des méthodes basées sur la « densité » des données répondant à ce cahier des charges ([Lelu & Ferhan 98], première variante de l'algorithme ultérieur GERMEN [Lelu et al. 06]).

7. Reconstitution des données et projections obliques

Notons $\mathbf{X}^{½}$ la matrice des données transformées : $x_{ti} \leftarrow \sqrt{x_{ti}}$ (T lignes = documents, I colonnes = mots)
Une fois convergé l'algorithme des KMA, soit $\mathbf{w}_t$ le vecteur projection du vecteur-ligne $\mathbf{x}^{½}_t$ sur l'ensemble des K axes de clusters $\mathbf{u}_k$ rassemblés dans la matrice **U** :

$$\mathbf{w}_t = \mathbf{x}^{½}_t \mathbf{U} \mathbf{D}$$

où **D** est la matrice diagonale des $\sqrt{\lambda^{(k)}}$
Ce qui peut s'écrire aussi, matriciellement :
$\mathbf{W} = \mathbf{X}^{½} \mathbf{U} \mathbf{D}$
Ou encore, en posant $\mathbf{V} = \mathbf{U} \mathbf{D}$:

$$\boxed{\mathbf{W} = \mathbf{X}^{½} \mathbf{V}}$$

On en déduit, dans le cas général où la matrice **V** n'est pas orthogonale ($\mathbf{V'\ V} \neq \mathbf{I}$), la reconstitution approximative au sens des moindres carrés :

$$\boxed{\mathbf{X}^{½} \cong \mathbf{W}\ \mathbf{V}^{+}}$$

où $\mathbf{V}^{+}$ est la pseudo-inverse de **V** :
$\mathbf{V}^{+} = (\mathbf{V'\ V})^{-1}\ \mathbf{V}'$ (dans le cas toujours vérifié en pratique où K<T)
A noter que si K = rang($\mathbf{X}^{½}$) alors $\mathbf{X}^{½} = \mathbf{W}\ \mathbf{V}^{+}$

La matrice des projections obliques $\mathbf{W_{obl}}$ se déduit de celle des projections orthogonales **W** :

$$\mathbf{W_{obl}} = \mathbf{W}\ (\mathbf{V'V})^{-1}$$

Ce qui permet d'écrire plus simplement la reconstitution :

$$\boxed{\mathbf{X}^{½} \cong \mathbf{W_{obl}}\ \mathbf{V}'}$$

$\mathbf{W_{obl}}$ s'écrit aussi, en fonction des facteurs normalisés **F** et **G** :

$$\mathbf{W_{obl}} = \mathbf{Dt}\ \mathbf{F}\ (\mathbf{G'\ Di\ G})^{-1}$$

Où **Dt** et **Di** sont les matrices diagonales des $\sqrt{x_{t.}}$ et des $x_{.i}$ respectivement.

Les projections obliques « augmentent le contraste » et constituent un type de représentation intermédiaire entre l'appartenance en tout ou rien à une seule classe d'une part, et les

projections orthogonales sur l'ensemble des axes des classes d'autre part. Alors que ces dernières sont nécessairement toutes positives, certaines valeurs de projections obliques peuvent être négatives, et peuvent s'interpréter en terme d'opposition entre les extrémités de tel ou tel « facteur local ».

Références

Arthur D. & Vassilvitskii S. (2006). How slow is the k-means method? *Proc. Of the 2006 Symposium on Computational Geometry* (SoCG).

Benzécri J.P. et coll. (1981). *Pratique de l'Analyse des Données : Linguistique et Lexicologie.* Dunod, Paris.

Domengès D., Volle M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, 35-1979 :3-84, Paris

Dumais S.T. (1994). Latent Semantic Indexing (LSI) and TREC-2. NIST special publication, N°500-215, pages 105-115, NIST.

Escofier B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionnelle. *Revue de Stat. Appliquée*, 26(4):29-37, Paris

Fichet B. et Gbegan A. (1985). Analyse factorielle des correspondances sur signes de présence-absence. In Diday et al. editors, *4e Journées Analyse des données et Informatique*. INRIA, Rocquencourt.

Kohonen T., Kaski S., Lagus K., Honkela T. (1995). Very large two-level SOM for the browsing of newsgroups. Proc. of *WWW'95 (5th International World Wide Web Conference)*, Paris.
<*http://websom.hut.fi/websom*>

Lebart L., Morineau A., Tabard N. (1977). *Techniques de la description statistique*. Dunod, Paris.

Lebart L., Rajman M. (2000). Computing Similarities. In Dale R., Moisl H., Somers H. editors : Handbook of Natural Language Processing, Marcel Dekker, pages 477-505, New York.

Lebart L., Salem A. (1994). *Statistique textuelle.* Dunod, Paris.

Lelu A., Tisseau-Pirot A.G. (1993). Emergence de catégories sémantiques à partir d'une base de résumés d'articles. In Anastex S.J. editor, *Proc. of JADT'98 (2emes Journées Internationales d'Analyse Statistique des données Textuelles),* pages 227-242, ENST, Paris.

Lelu A. (1993). Modèles neuronaux pour l'analyse de données documentaires et textuelles. Doctorat de l'université Paris 6, sous la direction de Ludovic Lebart.

Lelu A. (1994). Clusters *and* factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday, Y. Lechevallier & al. editors. *New Approaches in Classification and Data Analysis*, pages 241-248, Springer-Verlag, Berlin,

Lelu A., Tisseau-Pirot A.G., Adnani A. (1997). Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation. *Hypertextes et Hypermédias*, 1(1):23-55, Hermès, Paris,

Lelu A., Ferhan S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In Rizzi A. et al. editors, *Proc. of IFCS'98 (6th Conference of the International Federation of Classification Societies),* pages 206-209, Universita La Sapienza, Roma.

Lelu A., Halleb M., Delprat B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes. In Mellet S. editor, *Proc. of JADT'98 (4emes Journées Internationales d'Analyse Statistique des données Textuelles),* pages 391-400, UPRESA « Bases, Corpus et Langage », Université de Nice.

Lelu, A. (2003). Evaluation de trois mesures de similarité utilisées en sciences de l'information. *Information Sciences for Decision Making* 6 14–25.

Lelu A., Cuxac P., Cadot M. (2006) - Document stream clustering : an optimal and fine-grained incremental approach - *COLLNET'06* / International Workshop on Webometrics, Informetrics and Scientometrics, 10-12 mai 2006, Nancy

<*http://eprihtnts.rclis.org/archive/00006045/01/Collnet.pdf*>

Lelu A. (2006) – Clustering dynamique d'un flot de données : un algorithme incrémental et optimal de détection des maxima de densité – 8e Journées *EGC 2006* (Extraction et Gestion de Connaissances), Lille, 17-20 janvier 2006.

Lelu A., Cuxac P., Johansson J. (2006) - Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN - *JADT'06*, Besançon, 19-21 avril 2006.

Matusita K. (1955). Decision rules, based on the distance for problems of fit, two examples, and estimation. *Annals of Statistical Mathematics*, pages 631-640, Tokyo.

Plante P., Dumas L. et Plante A. (1997). *Atelier FX*. ATO, Département de Linguistique, Université du Québec à Montréal.

< *http://www.ling.uqam.ca/Ato/FX*>

Renyi A. (1966). *Calcul des probabilités.* Dunod, Paris.

Rhissassi H. and Lelu A. (1998). Indexation assistée et cartographie sémantique pour la génération automatique d'hypertexte. In Mojahid M. editors, *Proc. of CIDE'98*, pages 131-139, Europia Productions, INPT, Rabat, Maroc.

Salton G. (1968). *Automatic Information Organization and Retrieval*. Mac Graw Hill, NY.

Salton G. and Mac Gill M.J (1983). *Introduction to Modern Information Retrieval.* International Student Edition.

Zamir O., Etzioni O. (1999).Grouper : a dynamic Clustering Interface to Web Search Results. *Proc. of WWW'99 (8th International World Wide Web Conference).*