

Classification de variables

Classification autour de variables latentes

Ricco RAKOTOMALALA

PLAN

1. Classification automatique, typologie, etc.
2. Proximités, distances et composantes latentes
3. CAH de variables basée sur les composantes latentes
4. Interprétation des résultats
5. K-Means de variables
6. Une approche descendante originale (VARCLUS)
7. Complémentarité Classif. Individus & Classif. Variables
8. Références

La classification de variables

Classification automatique de variables

Qu'est-ce que la classification de variables ?

Créer des groupes de variables de manière à rassembler les variables qui portent les mêmes informations (redondantes, corrélées), et dissocier les variables qui expriment des informations complémentaires.

→ *C'est une autre manière de structurer les données. Une sorte d'analyse duale de la classification automatique des observations. Les deux analyses se complètent.*

Pourquoi la classification de variables ?

1. Comprendre les structures sous-jacentes qui organisent les données (oppositions, complémentarité, concomitance). On veut résumer l'information apportées par les données. Approche complémentaire de la classification des individus : mieux comprendre ce qui rassemble ou distingue les groupes.
2. Détecter les redondances (multi colinéarité) entre variables. Comprendre les principales dimensions que recèlent les données c.-à-d. décomposer l'information en unités élémentaires (plus ou moins) non redondantes. On peut déduire des variables synthétiques « résumé » à partir des groupes.
3. Réduction du nombre de variables. On peut l'utiliser comme un pré-traitement ou un post-traitement de la sélection de variables pour les autres techniques (ex. en apprentissage supervisé) c.-à-d.
 - a. Pour structurer l'espace de recherche lors de la sélection
 - b. Expliquer le positionnement des variables une fois la sélection réalisée.

Tableau de données Criminalité – USA en 1960

Variable	Statut	
CrimeRate	Illustrative	Crime rate: # of offenses reported to police per million population
Male14-24	Active	The number of males of age 14-24 per 1000 population
Southern	Active	Indicator variable for Southern states (0 = No, 1 = Yes)
Education	Active	Mean # of years of schooling x 10 for persons of age 25 or older
Expend60	Active	1960 per capita expenditure on police by state and local government
Expend59	Active	1959 per capita expenditure on police by state and local government
Labor	Active	Labor force participation rate per 1000 civilian urban males age 14-24
Male	Active	The number of males per 1000 females
PopSize	Active	State population size in hundred thousands
NonWhite	Active	The number of non-whites per 1000 population
Unemp14-24	Active	Unemployment rate of urban males per 1000 of age 14-24
Unemp35-39	Active	Unemployment rate of urban males per 1000 of age 35-39
FamIncome	Active	Median value of transferable goods and assets or family income in tens of \$
IncUnderMed	Active	The number of families per 1000 earning below 1/2 the median income

47 états

13 descripteurs, tous continus

1 variable illustrative « Crime rate »

Positionnement par rapport à l'ACP

Comprendre les redondances et les oppositions, n'est-ce pas déjà le rôle de l'ACP ?

ACP : Création de « facteurs » (dimensions) 2 à 2 orthogonaux. En associant les variables aux axes, on structure l'information portée par les données...

Mais...

(1) Difficulté à interpréter les axes. L'association « variable – axe » n'est pas toujours évidente (d'où l'utilisation des stratégies de rotation des axes d'ailleurs).

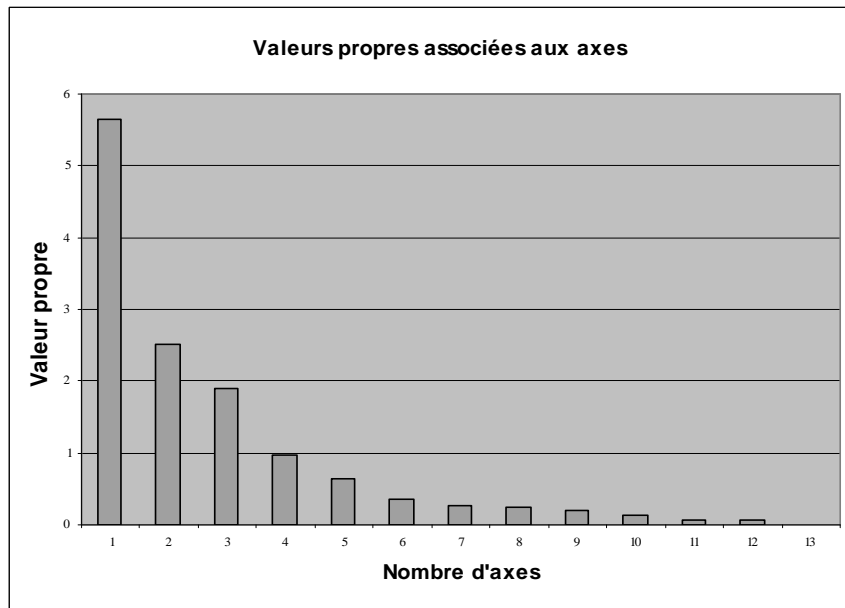
(2) Les variables sont dominées par une dimension principale qui « écrase » tout (cf. l'effet taille).

(3) Orthogonalité en cascade impose une contrainte (trop) forte dans la recherche des groupes de variables. Les axes suivants ne peuvent s'interpréter sans tenir compte des axes précédents (ex. Axe 2 : ACP sur les résidus de l'axe 1 avec la contrainte $\text{Axe 2} \perp \text{Axe 1}$)

Classification de variables : Peut être vue comme une technique proche de l'ACP où l'on aurait levé la contrainte d'orthogonalité entre les axes → ACP oblique.

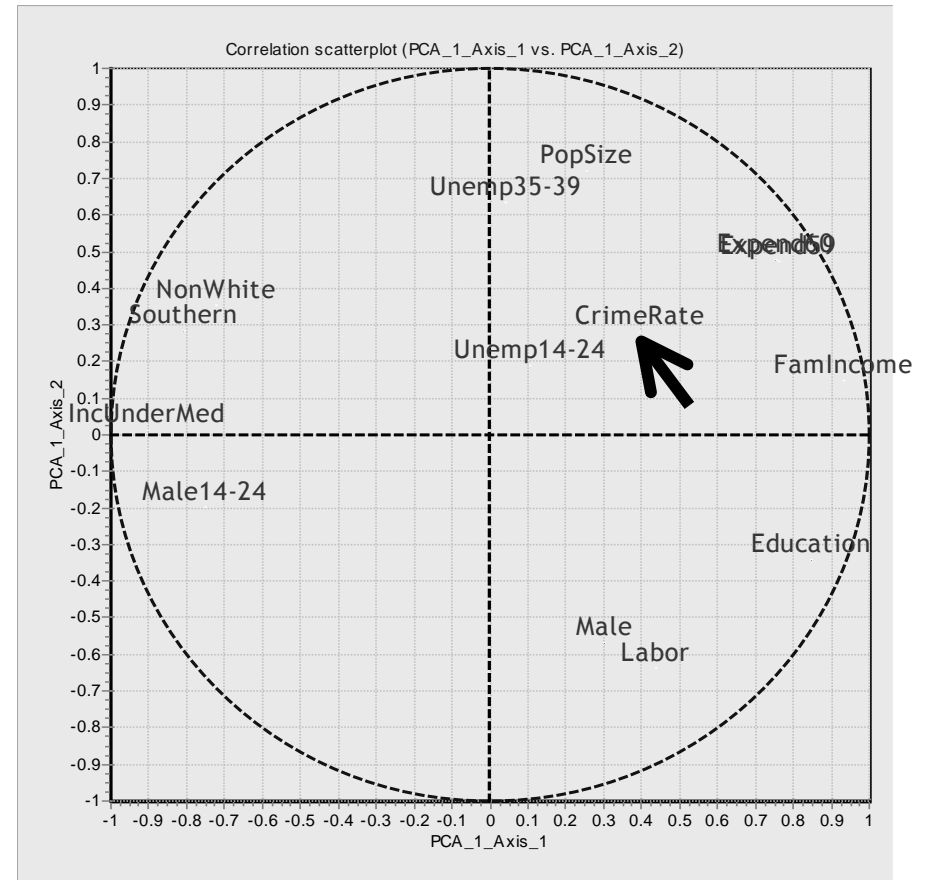
→ Classification autour des variables latentes : la première composante principale est associée à chaque groupe de variables.

ACP sur les données « Crime Dataset »



Attribute	Axis_1		Axis_2		Axis_3	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
-						
Male14-24	-0.754	57 % (57 %)	-0.1969	4 % (61 %)	0.1746	3 % (64 %)
Southern	-0.8083	65 % (65 %)	0.2848	8 % (73 %)	0.1359	2 % (75 %)
Education	0.8466	72 % (72 %)	-0.3404	12 % (83 %)	0.0408	0 % (83 %)
Expend60	0.7536	57 % (57 %)	0.476	23 % (79 %)	0.2433	6 % (85 %)
Expend59	0.7596	58 % (58 %)	0.4717	22 % (80 %)	0.2514	6 % (86 %)
Labor	0.4353	19 % (19 %)	-0.6365	41 % (59 %)	0.2087	4 % (64 %)
Male	0.3007	9 % (9 %)	-0.5691	32 % (41 %)	-0.4334	19 % (60 %)
PopSize	0.2551	7 % (7 %)	0.7203	52 % (58 %)	0.2712	7 % (66 %)
NonWhite	-0.7217	52 % (52 %)	0.3537	13 % (65 %)	0.2219	5 % (70 %)
Unemp14-24	0.1053	1 % (1 %)	0.1887	4 % (5 %)	-0.938	88 % (93 %)
Unemp35-39	0.0396	0 % (0 %)	0.6358	40 % (41 %)	-0.702	49 % (90 %)
FamIncome	0.9321	87 % (87 %)	0.15	2 % (89 %)	0.0742	1 % (90 %)
IncUnderMed	-0.9061	82 % (82 %)	0.0136	0 % (82 %)	-0.0226	0 % (82 %)
Var. Expl.	5.6518	43 % (43 %)	2.5202	19 % (63 %)	1.9059	15 % (78 %)

Cercle de corrélation avec la variable supplémentaire « Crime Rate »



Que faut-il comprendre ici ????

Proximités, distances et composantes latentes

Similarités (proximités) et distances entre variables

Distances entre groupes de variables

(1) Similarité r Coefficient de corrélation

$|r|$ ou r^2 Dans ce cas, le signe est ignoré, on veut une interprétation sous forme d'association et d'oppositions (comme en ACP)

(2) Distance $\sqrt{1-r}$ vs. $\sqrt{1-|r|}$ ou $\sqrt{1-r^2}$

(3) Distance entre une variable et un groupe de variables

- Saut minimum
- Saut maximum
- Moyenne des distances



Peut être étendu à la dissimilarité entre groupes de variables. On en déduit une stratégie d'agrégation pour la CAH par ex.




Critère de Ward si on peut définir une « variable moyenne » résumé d'un groupe de variable. La moyenne non pondérée n'est pas très satisfaisante car pas très interprétable.

Définition des composantes latentes

Idée : proposer un équivalent du barycentre pour les variables situées dans le groupe G_k

Z_k est défini de telle manière que $I_k = \sum_{j \in G_k} r^2(X_j, Z_k)$ soit minimum

$$\text{avec } Z_k = a_{1,k} X_{1,k} + a_{2,k} X_{2,k} + \dots$$

- (1) La variable « moyenne » (composante latente) est définie de manière à ce qu'elle soit le plus corrélée avec l'ensemble des variables du groupe
- (2) Avantage : Cette définition est totalement cohérente avec la définition des distances présentées précédemment. Voir aussi l'analogie avec la définition de la moyenne dans l'espace dual.
- (3) Z_k est le 1er axe factoriel de l'ACP sur les variables du groupe G_k 

Classification automatique comme un processus d'optimisation

K Nombre de groupes fixé

Inertie d'un groupe G_k

$$I_k = \sum_{j \in G_k} r^2(X_j, Z_k)$$

Inertie intra-classes

$$W = \sum_{k=1}^K \sum_{j \in G_k} r^2(X_j, Z_k)$$

W est le critère à optimiser (minimiser) lors du processus de classification

Théorème d'Huygens

$$T = B + W$$

Minimiser W = Maximiser B

Remarque : On retrouve les schémas de la classification automatique des individus. On peut donc appliquer les mêmes techniques (CAH, K-Means, etc.)



CAH de variables basée sur les composantes latentes

CAH de variables autour des composantes latentes VAR-CAH – Classification ascendante hiérarchique

Principe : Agréger au fur et à mesure les (groupes de) variables au sens de la minimisation de la perte d'inertie à chaque étape

Pour la fusion des groupes de variables G1 et G2 en G3,

$$\Delta = (\lambda_1 + \lambda_2) - \lambda_3$$

où λ est la valeur propre associée au 1er axe factoriel du groupe

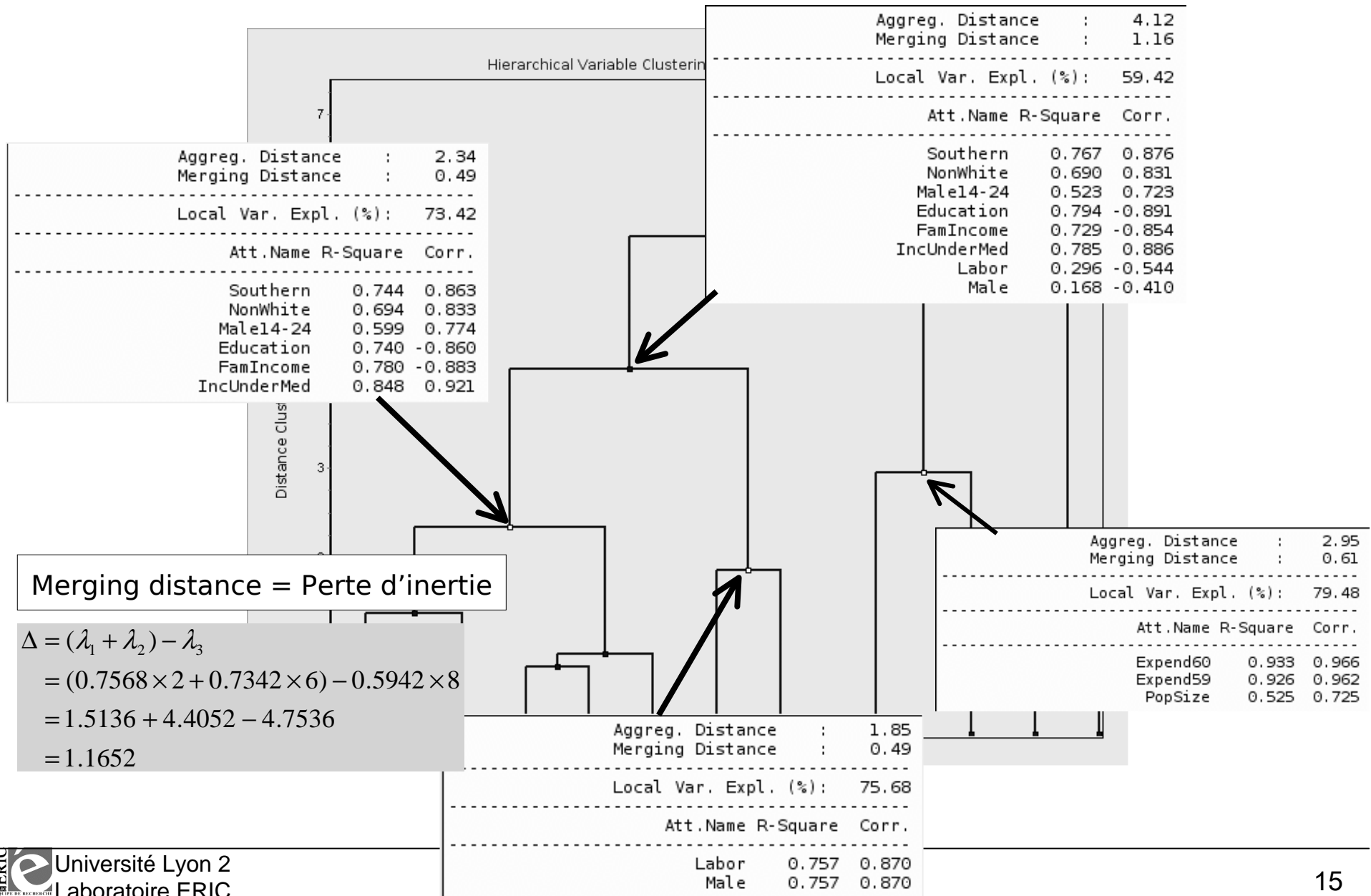
$$\Delta \geq 0 \text{ forcément}$$

- (1) On retrouve le processus « agglomératif » classique
- (2) A chaque regroupement, la variable moyenne est recalculée pour le groupe formé (Composante latente = 1er axe factoriel de l'ACP)
- (3) Pour chaque groupe, nous disposons de la corrélation de chaque variable avec la variable latente. Pour identifier les « parangons » des groupes.
- (4) Reste le problème classique de la classification : comprendre et interpréter les groupes !

VAR-CAH

Dendrogramme pour les données « Crime Dataset »
 Calcul de la perte d'inertie consécutive à une fusion

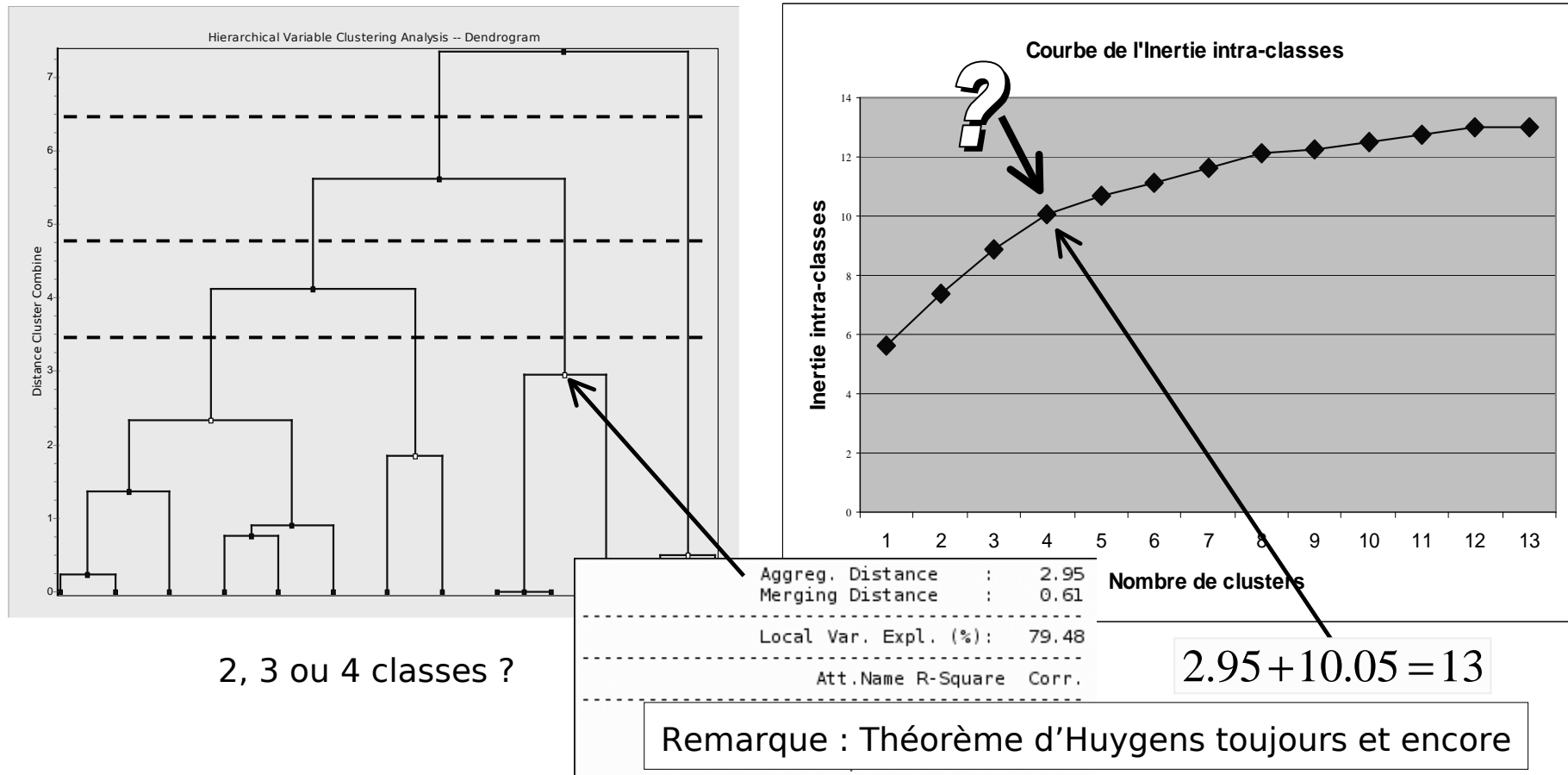
Hauteur du nœud = Hauteur du nœud de la fusion précédente + Perte d'inertie **4.12 = 2.95 + 1.16**



VAR-CAH

Détecter le nombre de groupes

Principe : Toujours la fameuse « loi du coude », signe d'une modification de la structure des données



C'est un problème sans fin. La meilleure manière de procéder est de guider le choix avec l'interprétation des résultats c.-à-d. utiliser les connaissances et les contraintes du domaine.

Lecture et Interprétation des résultats

Résultats - Description des classes

Variance expliquée par la
composante latente =
Val.Propre / Nb.Variables
Ex. 79.48% = 2.3843 / 3

Max r^2 avec les comp.
lat. des autres classes

$$ratio = \frac{1 - r_{own}^2}{1 - r_{next}^2}$$

#0, bon
>1, problème

Résumé des clusters

Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	3	2.3843	0.7948
3	6	4.4051	0.7342
4	2	1.5136	0.7568
Total		10.0489	0.7730

r^2 avec la composante
latente de sa propre
classe

Cluster members and R-square values

Cluster	Member	Own Cluster	Next Closest	1-R ² ratio
1	Unemp14-24	0.8730	0.0050	0.1277
	Unemp35-39	0.8730	0.0638	0.1357
2	Expend60	0.9334	0.3436	0.1015
	Expend59	0.9260	0.3569	0.1150
	PopSize	0.5249	0.0159	0.4827
3	Southern	0.7441	0.1011	0.2847
	NonWhite	0.6944	0.0213	0.3123
	Male14-24	0.5988	0.2473	0.5331
	Education	0.7396	0.1537	0.3076
	FamIncome	0.7798	0.5376	0.4762
	InclUnderMed	0.8485	0.3085	0.2191
4	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647

Nombre de
variables

% variance expliquée
par la classification

Valeur propre associée à
la composante latente

Degré d'adhésion à son groupe

Résultats – Structure des classes (Interprétation des classes)

Corrélation des variables avec les variables latentes de chaque classe

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Male14-24	1	-0.2511	-0.4973	0.7738	-0.1090
Southern	1	-0.0539	-0.3180	0.8626	-0.4714
Education	1	-0.1057	0.3920	-0.8600	0.5737
Expend60	1	0.0757	0.9661	-0.5862	0.0892
Expend59	1	0.0629	0.9623	-0.5974	0.0743
Labor	1	-0.3479	0.0546	-0.4169	0.8699
Male	1	0.1783	-0.1019	-0.2848	0.8699
PopSize	1	0.1243	0.7245	-0.1259	-0.3071
NonWhite	1	-0.0404	-0.1460	0.8333	-0.3842
Unemp14-24	1	0.9343	-0.0502	-0.1286	0.0704
Unemp35-39	1	0.9343	0.2255	0.0133	-0.2526
FamIncome	2	0.0733	0.7332	-0.8830	0.2726
IncUnderMed	1	-0.0258	-0.5554	0.9211	-0.2512

Remarques :

- (1) Une variable est associée à une classe, mais elle peut être corrélée à une autre classe : la contrainte d'orthogonalité est levée
- (2) Les classes s'interprètent en termes d'associations et oppositions (comme en ACP)
- (3) On comprend mieux le tableau d'adhésion au groupe

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R ² ratio
1	Unemp14-24	0.8730	0.0050	0.1277
	Unemp35-39	0.8730	0.0638	0.1357
2	Expend60	0.9334	0.3436	0.1015
	Expend59	0.9260	0.3569	0.1150
	PopSize	0.5249	0.0159	0.4827
3	Southern	0.7441	0.1011	0.2847
	NonWhite	0.6944	0.0213	0.3123
	Male14-24	0.5988	0.2473	0.5331
	Education	0.7396	0.1537	0.3076
	FamIncome	0.7798	0.5376	0.4762
4	IncUnderMed	0.8485	0.3085	0.2191
	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647

Seuil (arbitraire) fixé à $|r| = 0.7$ pour la mise en valeur des variables

Ex. FamIncome
 Own cluster (Cl.3)
 $-0.8830^2 = 0.7798$
 Next Closest (Cl.2)
 $0.7332^2 = 0.5376$

Résultats – Autres aides à l'interprétation

Corrélations entre les composantes latentes et traitement des variables supplémentaires

ACP oblique → les composantes latentes peuvent plus ou moins liées entre elles...

Results					
Y	X	r	r ²	t	Pr(> t)
VCHca_1_1	VCHca_1_2	0.0938	0.0088	0.6322	0.5305
VCHca_1_1	VCHca_1_3	-0.0617	0.0038	-0.4150	0.6801
VCHca_1_1	VCHca_1_4	-0.0975	0.0095	-0.6571	0.5145
VCHca_1_2	VCHca_1_3	-0.5169	0.2672	-4.0503	0.0002
VCHca_1_2	VCHca_1_4	-0.0272	0.0007	-0.1826	0.8559
VCHca_1_3	VCHca_1_4	-0.4033	0.1626	-2.9565	0.0049

Lien faible mais non négligeable entre
→ CL.2 et CL.3 (cf. justement FamIncome)
→ CL.3 et CL.4

Lien de « Crime Rate » avec les composantes latentes : essentiel pour l'interprétation

Results					
Y	X	r	r ²	t	Pr(> t)
CrimeRate	VCHca_1_1	0.0679	0.0046	0.4564	0.6503
CrimeRate	VCHca_1_2	0.6502	0.4228	5.7414	0.0000
CrimeRate	VCHca_1_3	-0.2162	0.0468	-1.4856	0.1443
CrimeRate	VCHca_1_4	0.2315	0.0536	1.5963	0.1174

États peuplés avec famille à hauts revenus, des dépenses pour la sécurité... emmène la criminalité ????

Quelle variable très importante
manque finalement dans cette étude



Méthodes des réallocations dynamiques

K-Means

K-Means - Maximiser (resp. minimiser) l'inertie inter-classes (resp. intra-classes) à K fixé

ALGORITHMME

Choisir K
 Définir au hasard K variables comme noyau des groupes
 Calculer les composantes latentes de chaque groupe
 TANT QUE non convergence
 POUR toutes les variables
 Affecter la variable à la composante latente qui lui est le plus proche (r^2)
 FIN POUR
 Calculer les composantes latentes de chaque groupe (Forgy)
 (cette étape peut être réalisée durant l'affectation ci-dessus – Mc Queen)
 FIN TANT QUE

Problèmes habituels
 Choix de K ?
 Pas de pistes pour solutions (K) alternatives ?

La lecture des résultats est identique



Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	4	3.1578	0.7895
2	4	3.1594	0.7899
3	2	1.7459	0.8730
4	3	1.7211	0.5737
Total		9.7843	0.7526

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R ² ratio
1	Male14-24	0.5339	0.4361	0.8265
	Expend60	0.8867	0.2297	0.1471
	Expend59	0.8938	0.2413	0.1400
	FamIncome	0.8435	0.6422	0.4375
2	Southern	0.8157	0.2961	0.2618
	Education	0.7776	0.3972	0.3689
	NonWhite	0.7629	0.1919	0.2934
	IncUnderMed	0.8031	0.6194	0.5173
3	Unemp14-24	0.8730	0.0013	0.1272
	Unemp35-39	0.8730	0.0358	0.1318
4	Labor	0.5444	0.2228	0.5862
	Male	0.7810	0.1224	0.2496
	PopSize	0.3957	0.0004	0.6045

Pour rappel, les résultats de la CAH

Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	3	2.3843	0.7948
3	6	4.4051	0.7342
4	2	1.5136	0.7568
Total		10.0489	0.7730

Il faut étudier en détail pour situer les différences

VARCLUS

Une approche descendante

Approche descendante

Quel intérêt ? S'arrêter dès que le partitionnement n'est plus pertinent

Algorithme

VARCLUS (L variables)

ACP avec les L variables

Rotation (QUARTIMAX) sur les 2 premiers axes

Si (Val.Propre 2ème axe fact. ≥ 1) Alors

Partition selon proximité (r^2) des variables avec les axes (L1 et L2)

VARCLUS (L1 variables)

VARCLUS (L2 variables)

Fin Si

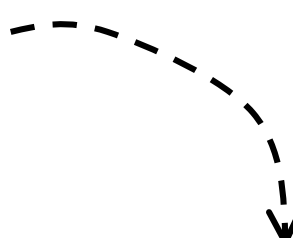
RETOUR

Remarques :

(1) Avantage : adaptée pour les grand nombre de variables

(2) Règle d'arrêt naturelle (modifiable)

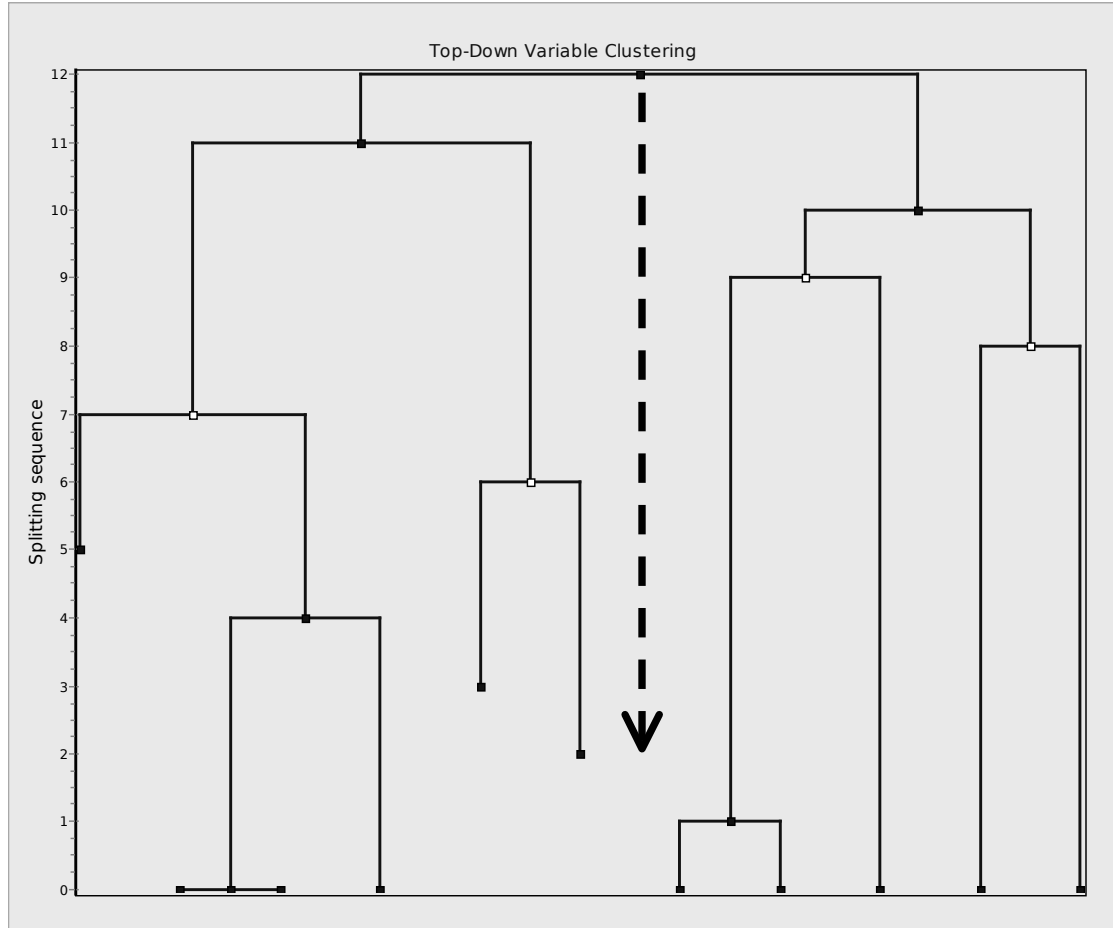
(3) La décroissance monotone de l'inertie intra n'est pas assurée (cf. artifice dans SAS – Réallocation à chaque étape)



Exemple : Base avec 52 variables et 3900 obs.
3 classes produites en ascendant et descendant
Ascendant # 5002 ms. ; Descendant # 797 ms.

VARCLUS - Résultats

La hauteur traduit seulement la séquence des segmentations. Pas d'interprétation en termes d'inertie.



Même résultat que la VARCAH

Cluster summary

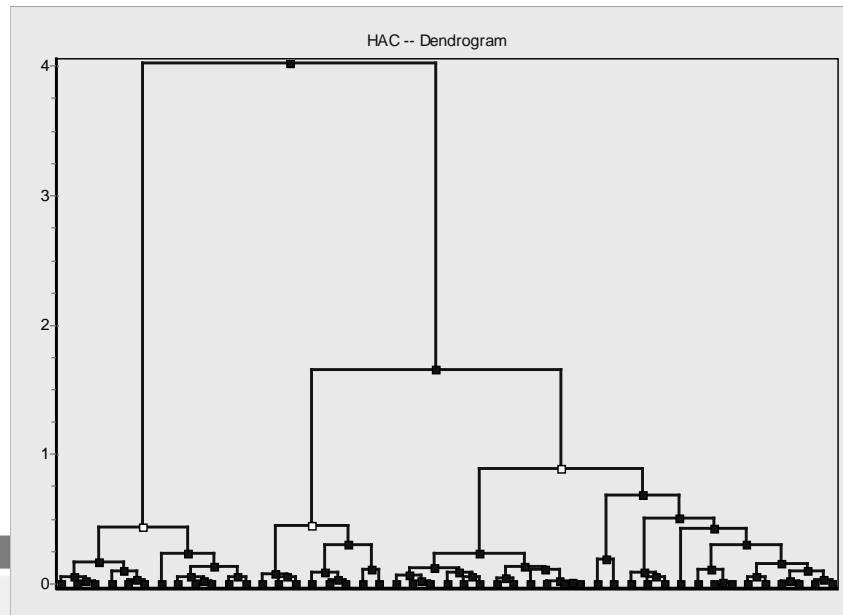
Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	2	1.5136	0.7568
3	6	4.4051	0.7342
4	3	2.3843	0.7948
Total		10.0489	0.7730

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R ² ratio
1	Unemp14-24	0.8730	0.0050	0.1277
	Unemp35-39	0.8730	0.0638	0.1357
2	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647
3	Male14-24	0.5988	0.2473	0.5331
	Southern	0.7441	0.1011	0.2847
	Education	0.7396	0.1537	0.3076
	NonWhite	0.6944	0.0213	0.3123
	FamIncome	0.7798	0.5376	0.4762
	IncUnderMed	0.8485	0.3085	0.2191
	IncOverMed	0.8485	0.3085	0.2191
4	Expend60	0.9334	0.3436	0.1015
	Expend59	0.9260	0.3569	0.1150
	PopSize	0.5249	0.0159	0.4827

Complémentarité avec la classification des individus

CAH sur les individus



La CAH propose « automatiquement » 3 groupes. On retrouve les mêmes structures : opposition nord-sud ; peuplé – haut revenu = dépenses de sécurité ; industriels + éducation vs. sud.

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[27.7 %] 13		Examples		[19.1 %] 9		Examples		[53.2 %] 25	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Southern	5.8	1.00 (0.00)	0.34 (0.48)	PopSize	4.7	91.22 (50.78)	36.62 (38.07)	Education	3.9	111.60 (7.92)	105.64 (11.19)
IncUnderMed	5.6	247.15 (14.87)	194.00 (39.90)	Expend59	4.2	116.11 (24.08)	80.23 (27.96)	Male	3.5	997.44 (31.00)	983.02 (29.47)
NonWhite	4.9	220.62 (115.07)	101.13 (102.83)	Expend60	4.2	122.78 (23.57)	85.00 (29.72)	Labor	3.1	578.72 (37.16)	561.19 (40.41)
Male14-24	4.7	152.77 (10.86)	138.57 (12.57)	FamIncome	3.8	636.56 (34.06)	525.38 (96.49)	FamIncome	1.5	545.40 (58.03)	525.38 (96.49)
Unemp35-39	0.3	34.62 (8.90)	33.98 (8.45)	Unemp35-39	2.3	39.78 (9.13)	33.98 (8.45)	Unemp14-24	0.4	96.48 (17.83)	95.47 (18.03)
PopSize	-0.1	35.31 (20.20)	36.62 (38.07)	Education	1.0	109.11 (6.01)	105.64 (11.19)	Expend59	-0.3	79.08 (21.89)	80.23 (27.96)
Unemp14-24	-0.8	91.85 (18.61)	95.47 (18.03)	Unemp14-24	0.4	97.89 (19.10)	95.47 (18.03)	Expend60	-0.3	83.76 (24.31)	85.00 (29.72)
Male	-2.0	968.85 (13.67)	983.02 (29.47)	NonWhite	-0.5	85.00 (40.74)	101.13 (102.83)	Male14-24	-1.7	135.56 (7.62)	138.57 (12.57)
Labor	-2.9	533.15 (39.32)	561.19 (40.41)	Labor	-0.7	553.00 (24.81)	561.19 (40.41)	Unemp35-39	-2.1	31.56 (7.10)	33.98 (8.45)
Expend60	-3.4	61.23 (12.17)	85.00 (29.72)	Southern	-1.6	0.11 (0.33)	0.34 (0.48)	IncUnderMed	-2.5	180.04 (24.21)	194.00 (39.90)
Expend59	-3.4	57.62 (11.39)	80.23 (27.96)	Male	-2.2	963.44 (20.53)	983.02 (29.47)	PopSize	-3.6	17.64 (14.84)	36.62 (38.07)
FamIncome	-5.0	409.92 (60.30)	525.38 (96.49)	IncUnderMed	-3.1	156.00 (15.39)	194.00 (39.90)	Southern	-3.9	0.08 (0.28)	0.34 (0.48)
Education	-5.2	91.77 (6.23)	105.64 (11.19)	Male14-24	-3.2	126.44 (5.96)	138.57 (12.57)	NonWhite	-4.0	44.80 (44.23)	101.13 (102.83)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

Bibliographie

En ligne

M. Qannari, E. Vigneau, P. Courcoux, « Une nouvelle distance entre variables. Application en classification », *Revue de Statistique Appliquée*, vol. 46, n°2, pp. 21-32, 1998.

http://archive.numdam.org/ARCHIVE/RSA/RSA_1998__46_2/RSA_1998__46_2_21_0/RSA_1998__46_2_21_0.pdf

E. Vigneau, M. Qannari, « Clustering of variables around latent components », in *Statistics, Simulation and Computation*, 32(4), pp.1131-1150, 2003.

http://www.nantes.inra.fr/les_recherches/sensometrie_et_chimiometrie/sensometrie/classification_de_variables

R. Rakotomalala, « Classification de variables », Tutoriels Tanagra pour le Data Mining.

<http://tutoriels-data-mining.blogspot.com/2008/03/classification-de-variables.html>

Ouvrages

J.P. Nakache, J. Confais - « Approche pragmatique de la classification », TECHNIP, 2004 ; pages 219 à 236.