Clustering variables

Clustering of variables around latent components

Ricco RAKOTOMALALA

Overview

1. Clustering variables
2. Correlations, distances and latent variables
3. HAC based on latent variables
4. Interpreting the results
5. K-Means algorithm for variables clustering
6. A top down approach: VARCLUS algorithm
7. Complementarity between clustering individuals and clustering variables
8. References

Clustering variables

What is the clustering variables?

The aim of the clustering variables is to detect subset of correlated variables. Thus, the variables which provide the same kind of information belong into the same group. The groups of variables reveal the main dimensionalities of the data. In a certain sense, it is more powerful than the factor analysis (e.g. principal component analysis) because it overcomes the orthogonality constraint between the factors.

→ This is another way to structure the data. A kind of dual analysis of the clustering individuals. Both analyses are complementary.

Why clustering variables?

1. Understand the underlying structures that organize the data. We want to summarize the information provided by the data. The approach is complementary to the clustering individuals.

2. Detect redundancies (collinearity) between variables. Understand key dimensions that contain the data i.e. decompose information into non-redundant interpretable basic units. We can deduce synthetic variables - similar to the factors from the factor analysis - from the groups.

3. Reduction of the number of variables. It can be used as a pre-treatment or post-treatment of the variables selection process for other techniques (e.g. in supervised learning) i.e.
   a. To structure the search space during the selection process
   b. To explain the real contribution of the variables once the selection is made.

« Crime » dataset – USA in 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 states (individuals)

13 quantitative variables

1 illustrative variable “Crime rate” (used for interpreting the results)

Do these techniques provide results of the same nature?

PCA: Creating orthogonal "components". By associating the variables to the components (e.g. the most correlated one), we can structure the variables…

But…

(1) We obtain often abstract components, that are difficult to interpret. In addition, a variable may be correlated to several components (the factor rotation strategies enable to overcome this drawback).

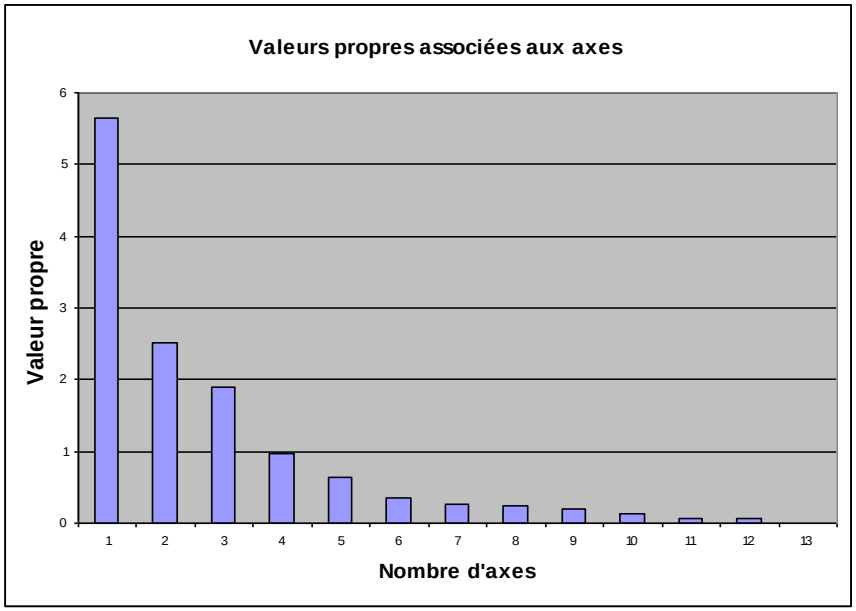
(2) Sometimes, a main dimension dominates, masking the other information (PCA on the partial correlation matrix may be a solution).

(3) The orthogonality constraint between the factors is sometimes too restrictive. In addition, a factor is interpretable only in accordance with the preceding factors.

→ Clustering variables may be viewed as a kind of oblique PCA

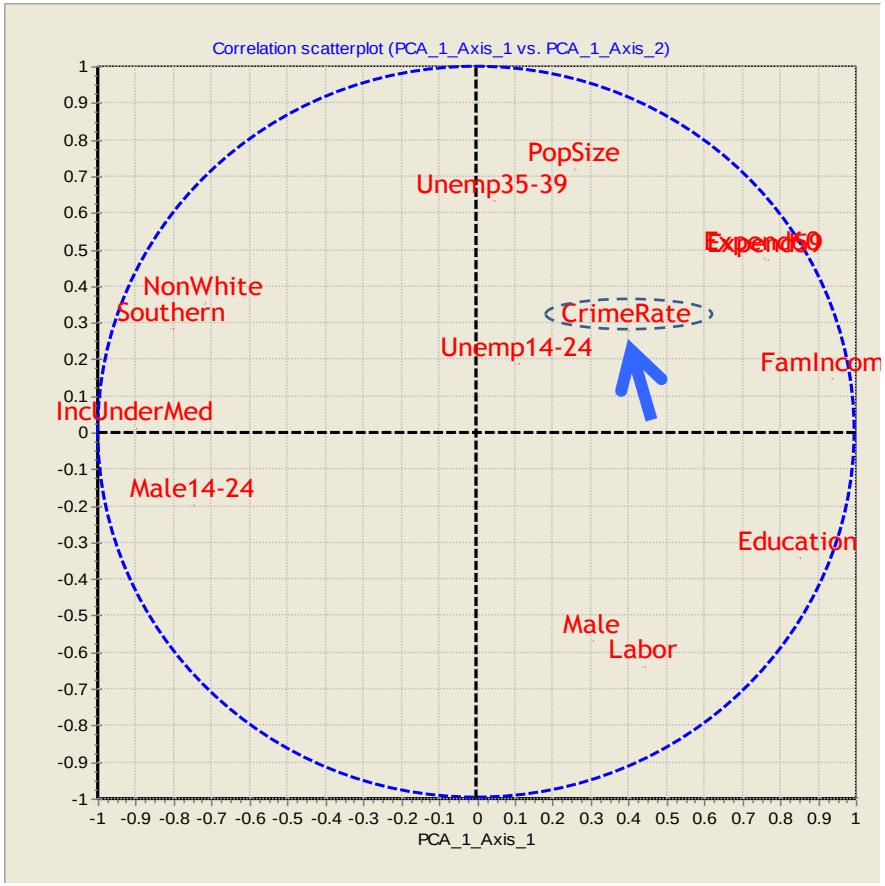
→ Clustering around latent component: the first component of the PCA is used to summarize a group of variables

PCA on the “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 %)*

Factor loading plot with the illustrative variable “crime rate”



Que faut-il comprendre ici ????

Correlation, distance and latent variable

(1) Similarity measure r Correlation coefficient

$|r|$ or r^2 In this second case, the sign of the relationship is ignored, we want an interpretation in the form of association and opposition (as in PCA)

(2) Dissimilarity: distance between variables

$\sqrt{1-r}$ vs. $\sqrt{1-|r|}$ or $\sqrt{1-r^2}$

We consider the direction and intensity of the relationship

We focus only on the intensity of the relationship

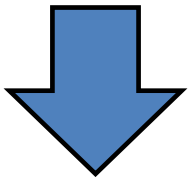
(3) Distance between subsets of variables (or between a variable and a set of variables)

→ Single linkage
→ Complete linkage
→ Average linkage
→ etc.

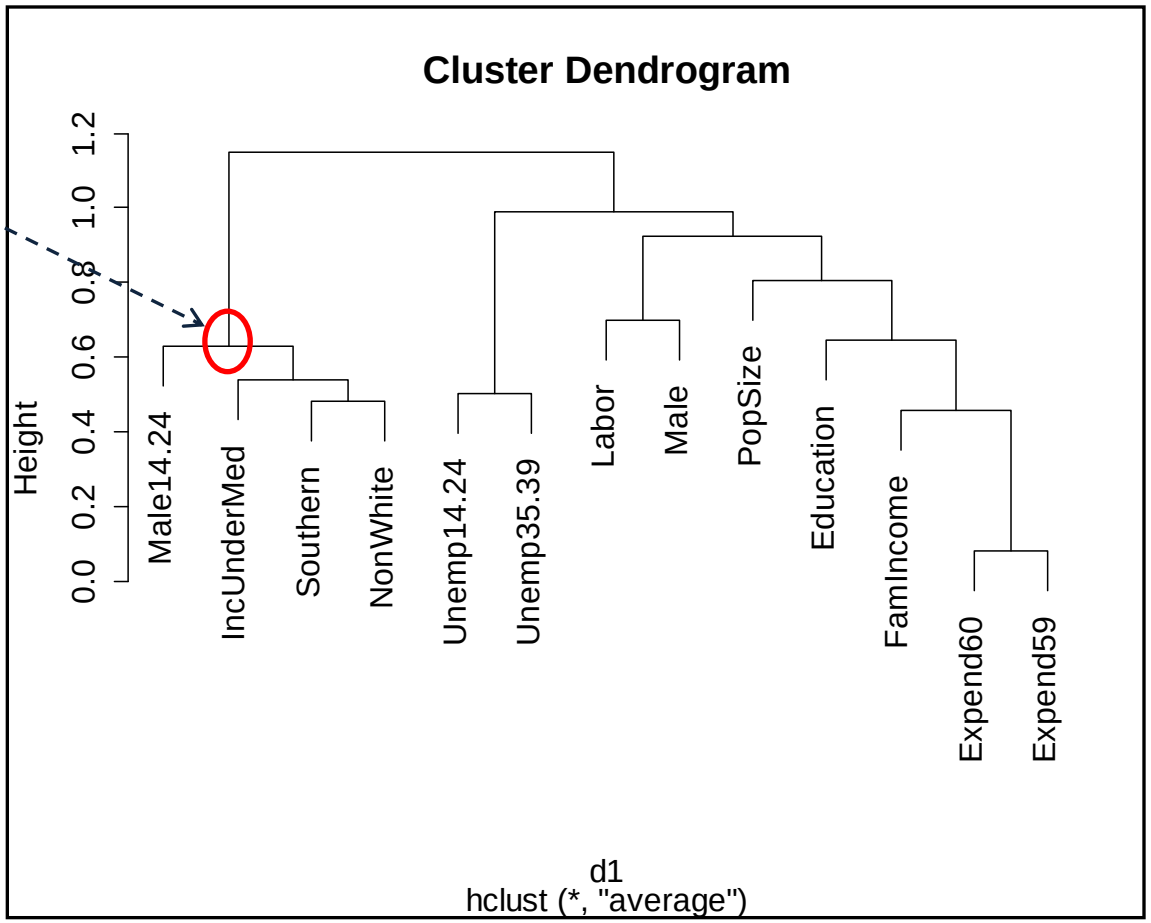
Aggregation strategy for HAC

Example under R: distance = SQRT(1 – r), Method = average link

Issue: How to set a "representative variable" which summarizes the information of 4 variables located on this node?



Idea of the latent component i.e. **the first component of the PCA**



Draw the parallel between this result and the factor loading plot of the PCA.

How to define a latent component?

Principle: Propose an equivalent of the "centroid" for a group G_k of variables

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

with $Z_k = a_{1,k} X_{1,k} + a_{2,k} X_{2,k} + \cdots$

(1) The "latent" variable is defined so that it is the most correlated (squared value) with the set of variables into the group

(2) The component can be viewed also as a synthetic variable which minimizes the sum of the squared distance ($d^2 = 1 - r^2$) to the existing variables. See the analogy with the definition of the centroid (mean) in the dual space.

(**3**) Z_k is the first principal component (from the PCA analysis) of the set of variables into G_k

Clustering algorithm as an optimization process

Using the latent components during the HAC process

K is the number of groups

Inertia for the group G_k

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

Within-group inertia (within-clusters variation)

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

W is the criterion to optimize (minimize) during the clustering process

Variation decomposition $T = B + W$

Minimize W ⇔ Maximize B

Note: There are the same schemes as clustering individuals. We can therefore apply the same techniques (HAC, K-Means, etc.)

HAC method for clustering variables

This approach can be extended to the processing of dataset with qualitative variables and/or mixed variables (quantitative / qualitative)

HAC for clustering of variables around latent components

VARHCA (into TANAGRA software) – Hierarchical agglomerative clustering

Principle : Step-by-step aggregation in the sense of the minimization of loss of inertia (variation)

For the merging of the groups G1 and G2 in G3…

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

where λ is the eigenvalue related to the first principal component of the group

$\Delta \geq 0$ necessarily

(1) We find the standard "agglomeration" process

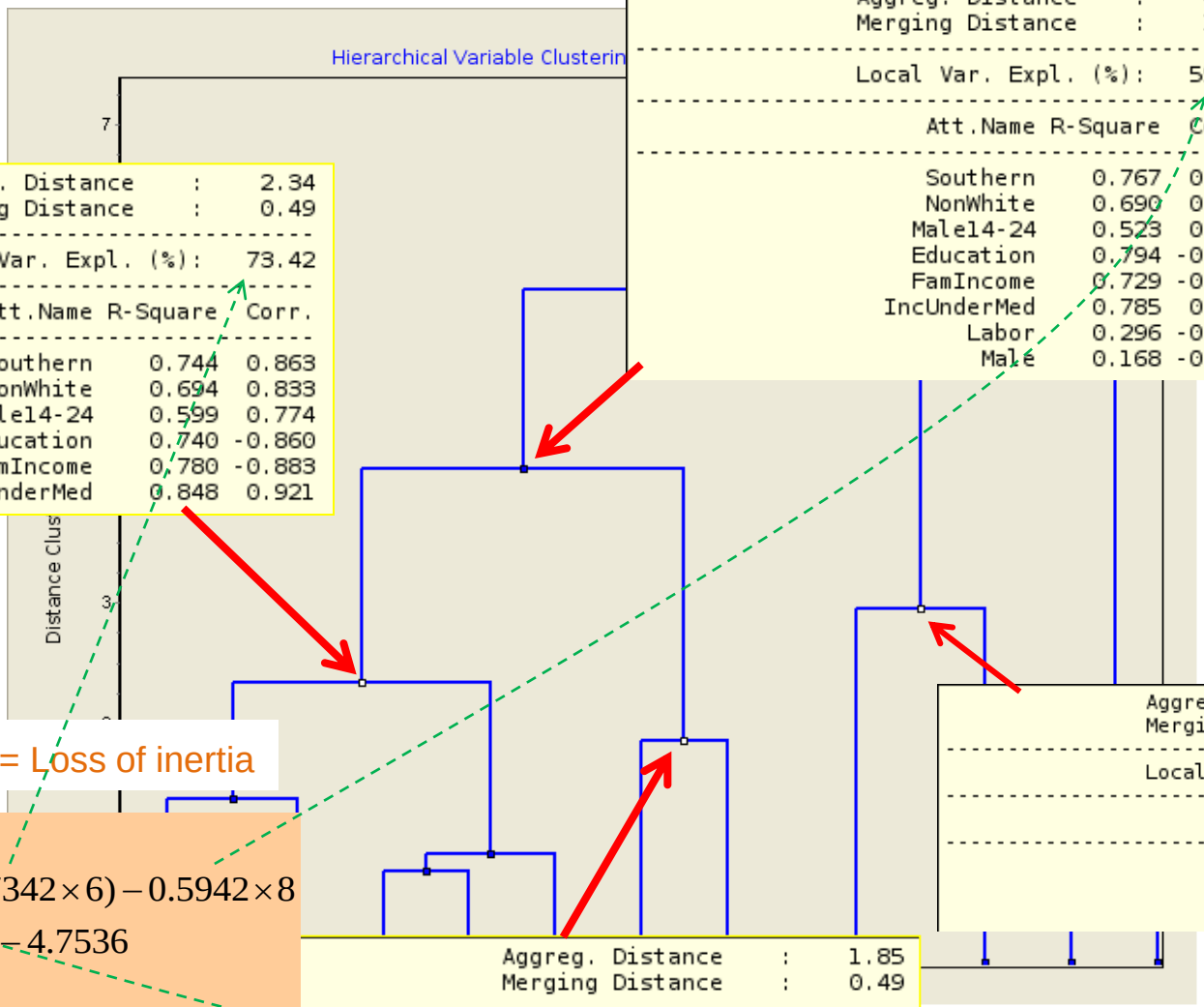
(2) At each step, the first component of PCA is computed for the new subset of variables

(3) For each group, we calculate the correlation of each variable with the latent component, in order to identify the most representative variables of the group.

(4) The usual problem of the clustering process remains: understanding and interpreting groups!

Height of the node = height of the preceding merging + loss of inertia

$$4.12 = 2.95 + 1.16$$



Aggreg. Distance	4.12
Merging Distance	1.16
Local Var. Expl. (%)	59.42

Att.Name	R-Square	Corr.
Southern	0.767	0.876
NonWhite	0.690	0.831
Male14-24	0.523	0.723
Education	0.794	-0.891
FamIncome	0.729	-0.854
IncUnderMed	0.785	0.886
Labor	0.296	-0.544
Male	0.168	-0.410

Aggreg. Distance	2.34
Merging Distance	0.49
Local Var. Expl. (%)	73.42

Att.Name	R-Square	Corr.
Southern	0.744	0.863
NonWhite	0.694	0.833
Male14-24	0.599	0.774
Education	0.740	-0.860
FamIncome	0.780	-0.883
IncUnderMed	0.848	0.921

Aggreg. Distance	2.95
Merging Distance	0.61
Local Var. Expl. (%)	79.48

Att.Name	R-Square	Corr.
Expend60	0.933	0.966
Expend59	0.926	0.962
PopSize	0.525	0.725

Merging distance = Loss of inertia

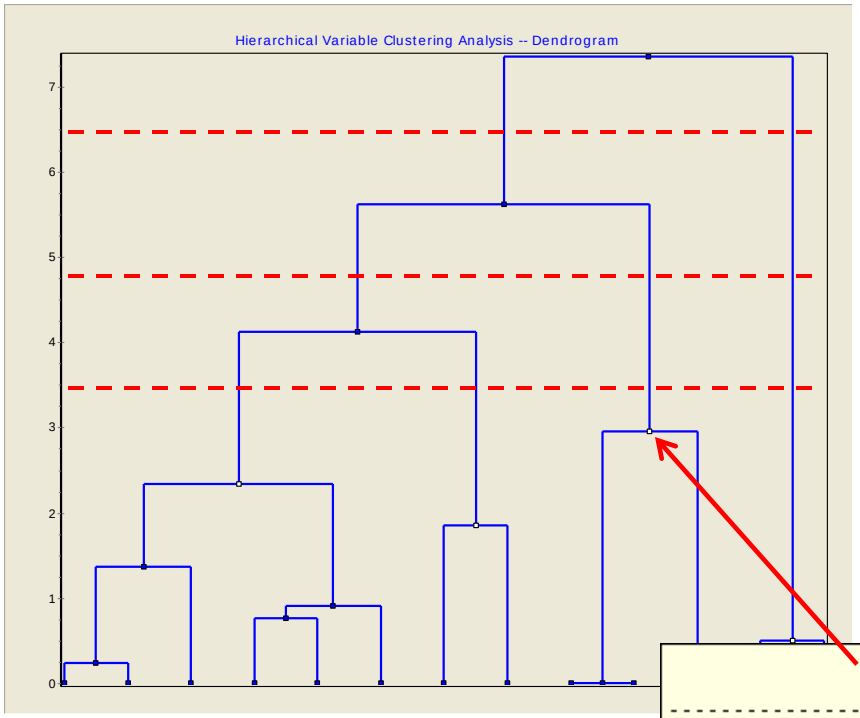
$$\begin{aligned}\Delta &= (\lambda_1 + \lambda_2) - \lambda_3 \\ &= (0.7568 \times 2 + 0.7342 \times 6) - 0.5942 \times 8 \\ &= 1.5136 + 4.4052 - 4.7536 \\ &= 1.1652\end{aligned}$$

Aggreg. Distance	1.85
Merging Distance	0.49
Local Var. Expl. (%)	75.68

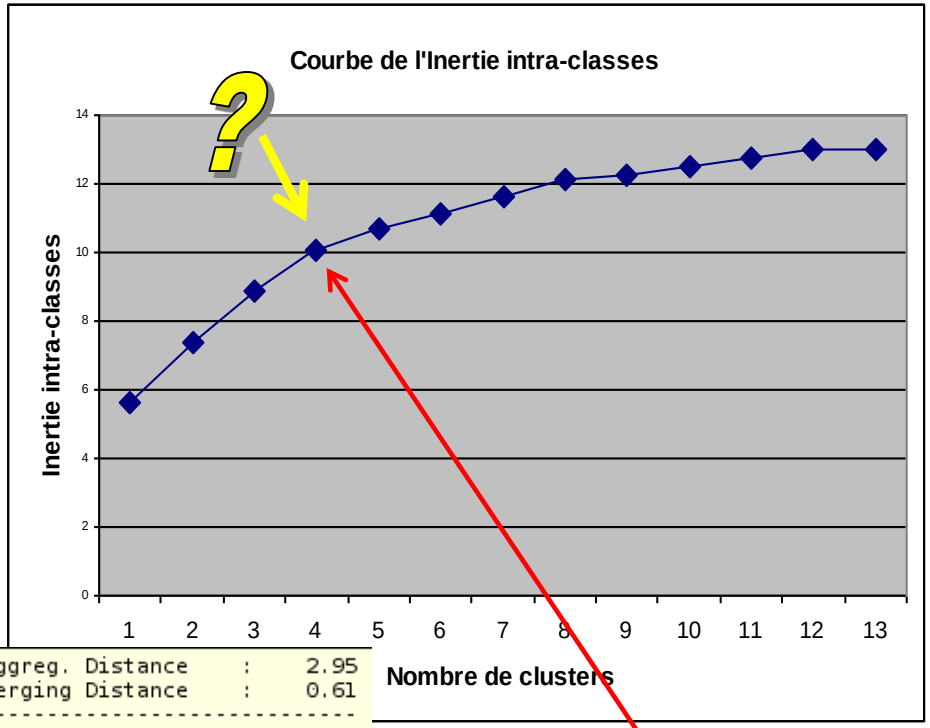
Att.Name	R-Square	Corr.
Labor	0.757	0.870
Male	0.757	0.870

Determining the number of clusters

Principle: The “elbow” method. Identifying a significant modification of the data structure.



2, 3 or 4 clusters?



Aggreg. Distance :	2.95
Merging Distance :	0.61
Local Var. Expl. (%):	79.48
Att.Name R-Square	Corr.

Note: Decomposition of the variation

$2.95 + 10.05 = 13$

This remains an open issue. The results must be corroborated by the domain expertise.

Reading and interpreting the results

Variance explained by the latent component
= Eigenvalue / Number of Variables
Ex. 79.48% = 2.3843 / 3

Max r² with the latent component of the other clusters

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

#0, good
>1, bad

r² with the latent component of the own group

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

Number of variables

Eigenvalue related to the latent component

% of variance explained by the clustering process

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² 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
	IncUnderMed	0.8485	0.3085	0.2191
4	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647

Quality of the membership of a variable to its group

Correlation of the variable with the latent component of each group.

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

An arbitrary threshold is used |r| = 0.7 to highlight the correlations into Tanagra

Ex. FamIncome

Own cluster (Cl.3)

$-0.8830^2 = 0.7798$

Next Closest (Cl.2)

$0.7332^2 = 0.5376$

Note:

(1) A variable may be a membership to a group, but highly correlated to another group. We must detect this kind of inconsistency.

(2) For each group, the correlations to the latent variable may be positive or negative.

(3) The "cluster correlations" table completes the "cluster members" table, and it gives a enhanced explanation of the layout of the clusters.

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² 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
	IncUnderMed	0.8485	0.3085	0.2191
4	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647

The latent components may be more or less correlated each other.

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

A weak relationship, but significant, between CL.2 and CL.3 (see the influence of FamIncome); and between CL.3 and CL.4.

Relationship with the "crime rate" variable. It seems that it is highly related to the CL.2.

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

The states with high family income, high expenditure to security... are the states where the criminality is high? This seems really surprising.

A very important variable misses in this study. Which one?

K-Means clustering around latent components

ALGORITHM

Set K (the number of groups)
Choose randomly K variables as latent variable for each group
This variable is the first latent component of the group
DO WHILE no convergence
  FOR EACH variable
    Assign the variable to the closest latent component (r^2)
  END FOR
 Update the latent component for each group (Forgy)
 (this step may be realized for each allocation above - Mc Queen)
END DO

Typical issues for K-Means:

How to choose K?

The reading of the results is the same than for HAC

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

As a reminder, the following are the results for the HAC

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

We need to study in detail to understand differences

VARCLUS

A top-down approach

Top-down approach for clustering variables

Advantage? Quickness. We can stop as soon as a subdivision is not relevant.

Algorithm

VARCLUS (L variables)
  PCA with the L variables
  Rotation (QUARTIMAX) on the 2 first components
  IF (Eigenvalue of the 2nd component ≥ 1) THEN
    Subdivision according to " r^2 " of the variables with the components (L1 and L2)
    VARCLUS (L1 variables)
    VARCLUS (L2 variables)
 END IF
RETURN

E.g. : Dataset with 52 variables and 3900 obs. We want to obtain 3 groups.
Bottom-up (HAC) # 5002 ms. ; Top-down # 797 ms.

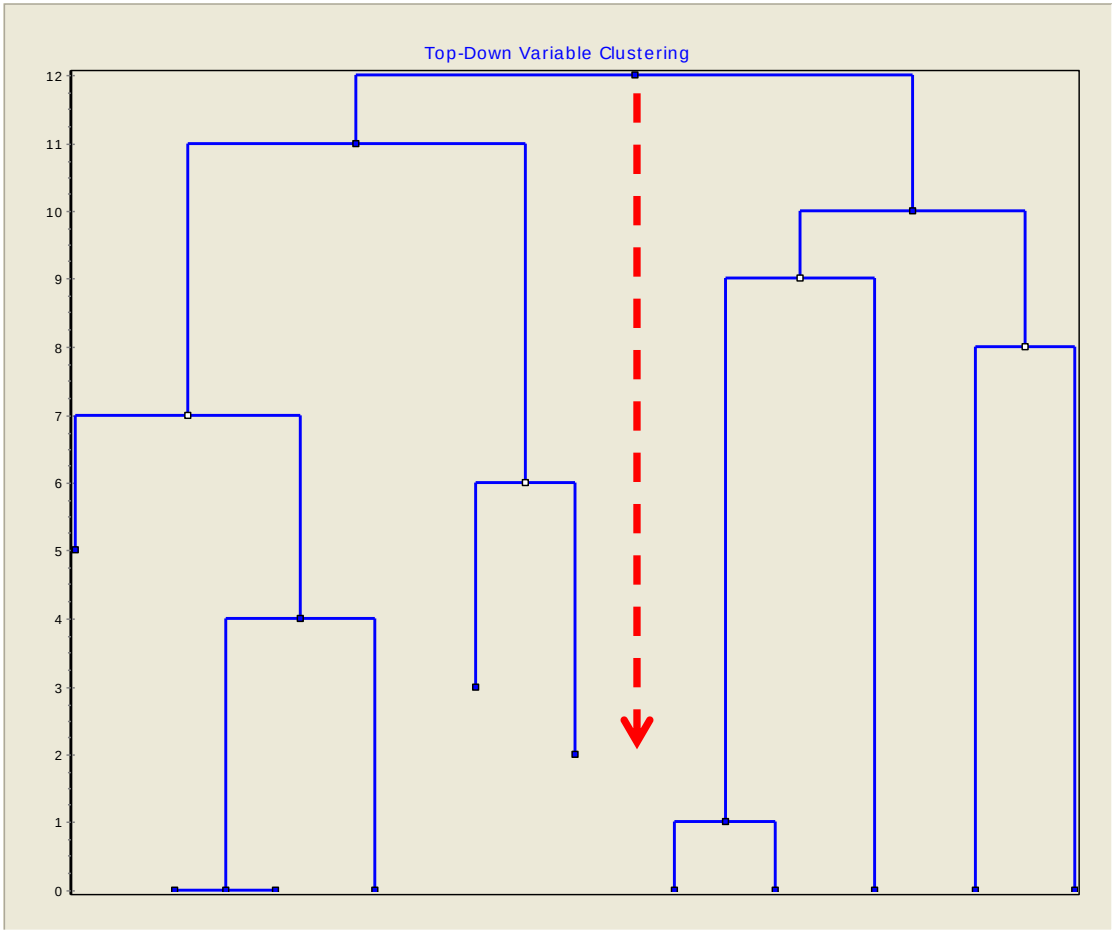
Note:

(1) Main advantage: Fast processing of large dataset

(2) Understandable stopping rule (that we can modify)

(3) The decreasing monotonically of the within-group inertia is not guaranteed (some tools reassigns the variables to the clusters at each step to minimize the within-group inertia).

The height shows only the sequence of subdivisions.



Same results as HAC

Cluster summary

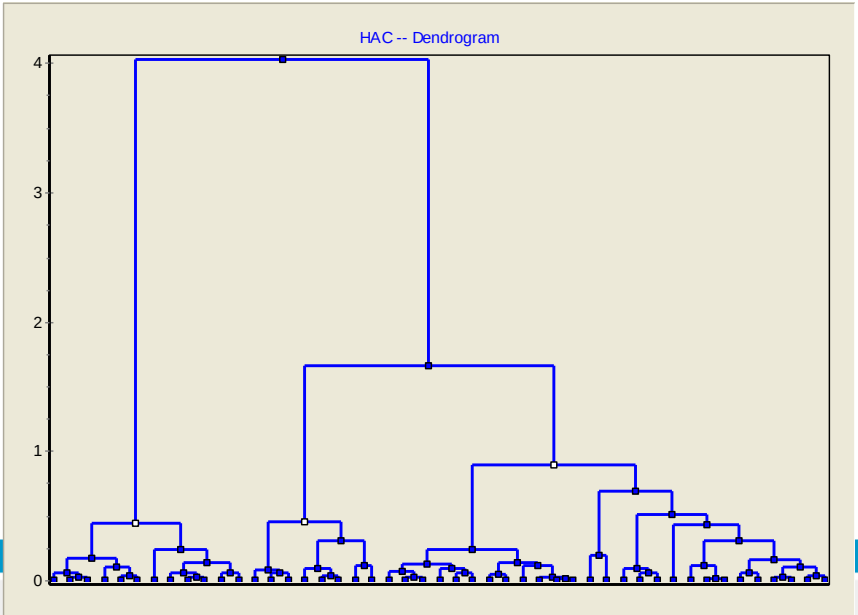
Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	2	1.5136	0.7568
3	6	4.4051	0.7342
4	3	2.3843	0.7948
Total		10.0489	0.7730

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
4	Expend60	0.9334	0.3436	0.1015
	Expend59	0.9260	0.3569	0.1150
	PopSize	0.5249	0.0159	0.4827

Complementarity with clustering individuals

HAC on the individuals (Crime dataset)



The HAC leads to a subdivision in 3 groups. We find the same interpretations: north vs. south; populated + high income = security expenditures; industrious + high level of education vs. south.

Cluster_HAC_1=c_hac_1			
Examples			[27.7 %] 13
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Southern	5.8	1.00 (0.00)	0.34 (0.48)
IncUnderMed	5.6	247.15 (14.87)	194.00 (39.90)
NonWhite	4.9	220.62 (115.07)	101.13 (102.83)
Male14-24	4.7	152.77 (10.86)	138.57 (12.57)
Unemp35-39	0.3	34.62 (8.90)	33.98 (8.45)
PopSize	-0.1	35.31 (20.20)	36.62 (38.07)
Unemp14-24	-0.8	91.85 (18.61)	95.47 (18.03)
Male	-2.0	968.85 (13.67)	983.02 (29.47)
Labor	-2.9	533.15 (39.32)	561.19 (40.41)
Expend60	-3.4	61.23 (12.17)	85.00 (29.72)
Expend59	-3.4	57.62 (11.39)	80.23 (27.96)
FamIncome	-5.0	409.92 (60.30)	525.38 (96.49)
Education	-5.2	91.77 (6.23)	105.64 (11.19)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_2			
Examples			[19.1 %] 9
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
PopSize	4.7	91.22 (50.78)	36.62 (38.07)
Expend59	4.2	116.11 (24.08)	80.23 (27.96)
Expend60	4.2	122.78 (23.57)	85.00 (29.72)
FamIncome	3.8	636.56 (34.06)	525.38 (96.49)
Unemp35-39	2.3	39.78 (9.13)	33.98 (8.45)
Education	1.0	109.11 (6.01)	105.64 (11.19)
Unemp14-24	0.4	97.89 (19.10)	95.47 (18.03)
NonWhite	-0.5	85.00 (40.74)	101.13 (102.83)
Labor	-0.7	553.00 (24.81)	561.19 (40.41)
Southern	-1.6	0.11 (0.33)	0.34 (0.48)
Male	-2.2	963.44 (20.53)	983.02 (29.47)
IncUnderMed	-3.1	156.00 (15.39)	194.00 (39.90)
Male14-24	-3.2	126.44 (5.96)	138.57 (12.57)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_3			
Examples			[53.2 %] 25
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Education	3.9	111.60 (7.92)	105.64 (11.19)
Male	3.5	997.44 (31.00)	983.02 (29.47)
Labor	3.1	578.72 (37.16)	561.19 (40.41)
FamIncome	1.5	545.40 (58.03)	525.38 (96.49)
Unemp14-24	0.4	96.48 (17.83)	95.47 (18.03)
Expend59	-0.3	79.08 (21.89)	80.23 (27.96)
Expend60	-0.3	83.76 (24.31)	85.00 (29.72)
Male14-24	-1.7	135.56 (7.62)	138.57 (12.57)
Unemp35-39	-2.1	31.56 (7.10)	33.98 (8.45)
IncUnderMed	-2.5	180.04 (24.21)	194.00 (39.90)
PopSize	-3.6	17.64 (14.84)	36.62 (38.07)
Southern	-3.9	0.08 (0.28)	0.34 (0.48)
NonWhite	-4.0	44.80 (44.23)	101.13 (102.83)
Discrete attributes : [Recall] Accuracy			

References

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

J. Saracco, M. Chavent, V. Kuentz, "Clustering of categorical variables around latent variables", Cahiers du GTREThA, N°2010-02, 2010.

Tanagra Tutorials, "Variable clustering (VARCLUS)", 2008.

SAS/STAT® 9.3 User's Guide, "The VARCLUS Procedure".

I. Endrizzi, "Clustering of variables around latent components: an application in consumer science", Universita di Bologna, 2008.