From Cluster Analysis to Classification

Dominique Joye

University of Lausanne

10th of December 2021

Introduction

Cluster analysis is a statistical techniques very often used, at least since the development of modern computers and statistical software more than fifty years ago.
It is important to have a look to these techniques for three important reasons at least:

- The ease of interpretation or, even more, of communication of results by comparison to other multivariate techniques.
- The technical discussion behind the choice of a particular technique.
- The choice of variables on which the analysis is based and the way they are measured

Classification and information reduction

In survey analysis, information reduction is a very important question: we know that answers to a single item has a lot of variability, also function of external causes: time of the day, contextual influences, etc.
The most frequent tradition of analysis is the use of latent variables as a way to summarize information. In such a case, there is at least 2 perspectives

- exploratory with techniques like principal component analysis
- confirmatory with the use of Confirmatory factor analysis, particularly in the line of SEM

Most often cluster analysis is an exploratory technique. More confirmatory approaches are possible but out of the scope of this presentation.

Back to measurement theory?

Even if the traditional discussion between nominal, ordinal and interval variables is not really simple and unambiguous,

- PCA or factor analysis summarize a set of variables by creating a group of latent, interval variables
- Cluster analysis create a categorical variable as the summary

By comparison, in the analysis of social position, we have the same: a socio-economic index is an interval variable when a class schema is a nominal variable.
Of course the theoretical dimensions underlying a class schema are important for interpretation!

Wright schema of classes, from the book *Classes*

Owners		Non-owners			Org. Assets
Bourgeoisie		Expert Managers	Semi Credentialled Managers	Un-Credentialled Managers	+
Small Employers		Experts Supervisors	Semi Credentialled Supervisors	Un-Credentialled Supervisors	=
Petty Bourgeoisie		Experts non-managers	Semi Credentialled Workers	Proletarians	-
		Skil +	Skil =	Skil -	

Distances

The definitions of groups imply to have a measure of distance or proximity between objects. Most often we use the euclidean distance

$$d = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (1)$$

But many other distances are possible, even similarity measures like correlations.

Single linkage

$$A5 = \begin{pmatrix} 0 & 2 & 3 & 7 & 9 \\ 2 & 0 & 4 & 6 & 5 \\ 3 & 4 & 0 & 4 & 5 \\ 7 & 6 & 4 & 0 & 3 \\ 9 & 5 & 5 & 3 & 0 \end{pmatrix} \quad A4 = \begin{pmatrix} 0 & 3 & 6 & 5 \\ 3 & 0 & 4 & 5 \\ 6 & 4 & 0 & 3 \\ 5 & 5 & 3 & 0 \end{pmatrix} \tag{2}$$

$$A3 = \begin{pmatrix} 0 & 4 & 5 \\ 4 & 0 & 3 \\ 5 & 3 & 0 \end{pmatrix} \quad A2 = \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix} \tag{3}$$

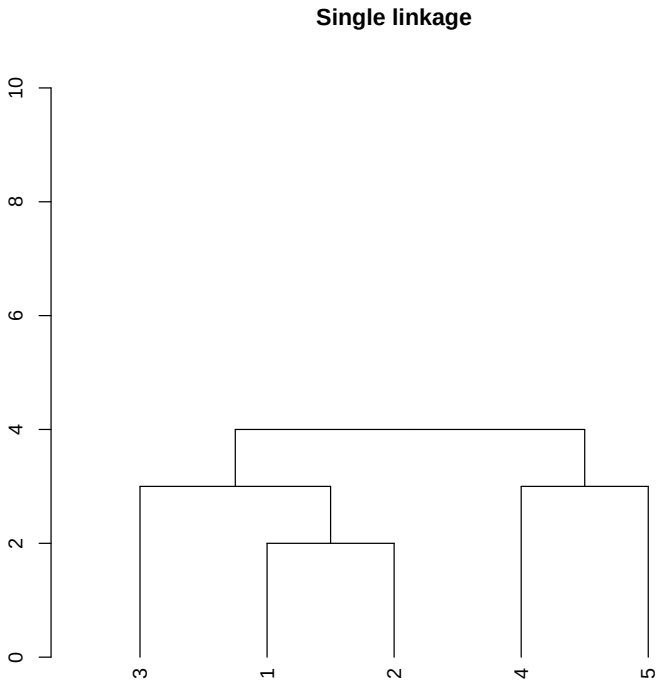
L=2 (1-2)
L=3 ((1-2)-3)
L=3 (4-5)
L=4 (((1-2)-3)-(4-5))

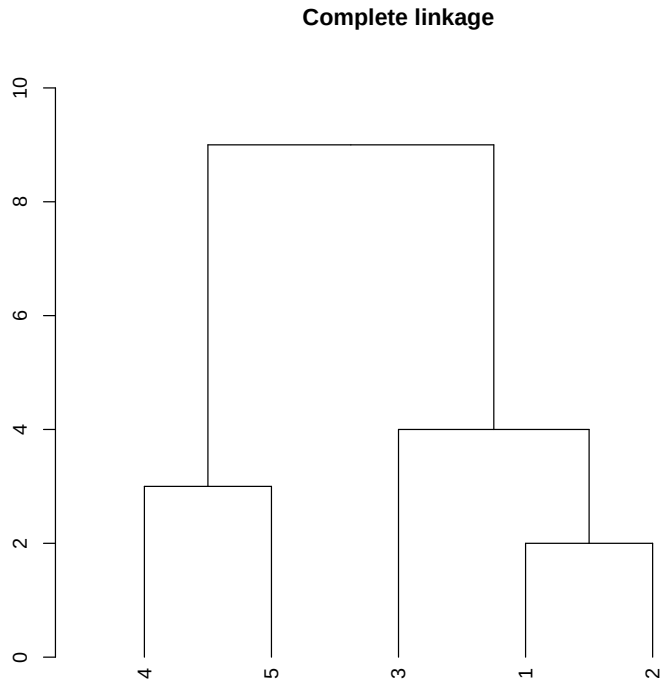
Complete linkage

$$A5 = \begin{pmatrix} 0 & 2 & 3 & 7 & 9 \\ 2 & 0 & 4 & 6 & 5 \\ 3 & 4 & 0 & 4 & 5 \\ 7 & 6 & 4 & 0 & 3 \\ 9 & 5 & 5 & 3 & 0 \end{pmatrix} \quad A4 = \begin{pmatrix} 0 & 4 & 7 & 9 \\ 4 & 0 & 4 & 5 \\ 7 & 4 & 0 & 3 \\ 9 & 5 & 3 & 0 \end{pmatrix} \tag{4}$$

$$A3 = \begin{pmatrix} 0 & 4 & 9 \\ 4 & 0 & 5 \\ 9 & 5 & 0 \end{pmatrix} \quad A2 = \begin{pmatrix} 0 & 9 \\ 9 & 0 \end{pmatrix} \tag{5}$$

L=2 (1-2)
L=3 (4-5)
L=4 ((1-2)-3)
L=9 (((1-2)-3)-(4-5))


Single linkage
0
2
4
6
8
10
3
1
2
4
5


Complete linkage
0
2
4
6
8
10
4
5
3
1
2

And other techniques...

Of course, between these extremes, we can imagine

- Mean, weighted or not
- Median, at the risk of reversal
- Ward... or a variance criterium (2 units are combined with the minimum increase in error sum of squares)
- Lance & Williams, flexible method (available in the cluster package, "agnes" routine)

In fact, the Ward method seems very often used in many publications.

and some examples

See the additional pdf document.
These examples were based on [Everitt(1974)].

Hierarchical, in summary

- Advantages
  - Robust in some cases or with some methods
  - Easy to read
- Problems
  - Optimize the path, not the classification
  - Even with modern computers, some limitations for big datasets

In summary, one piece of the puzzle but not a magic tool

Kmeans or partitioning techniques

From a very long time, a strategy in 3 steps, for a given number of clusters

- To find a start
- To allocate the units to the closest group
- To reallocate according the changes

Problems (and solutions?) for Kmeans

- How to find start? A hierarchical analysis could help
- A distance criteria could tend to spherical groups: robust alternative proposed by Kaufman and Rousseuw
- How to find the good number of groups to analyse?

In summary

- Mix of hierarchical and non-hierarchical can help in any case
- And the stability of a solution is also an indication.

ISSP 2007 Cultural behavior

The ISSP 2007 called leisure and sport, had the following question: *How often do you do each of the following activities in your free time?*

- 1. Daily
- 2. Several times a week
- 3. Several times a month
- 4. Several times a year or less often
- 5. Never

The choice was between

- a. Watch TV, DVD, videos
- b. Go to the movies
- c. Go out shopping
- d. Read books
- e. Attend cultural events such as concerts, live theatre...
- f. Get together with relatives
- g. Get together with friends
- h. Play cards or board games
- i. Listen to music
- j. Take part in physical activities such as sports... gym ... walk
- k. Attend sporting events as a spectator,
- l. Do handicrafts such as needle work, wood work, etc.
- m. Spend time on the Internet/PC

Variables scaling

The choice of measurement is of course crucial. We can have 2 strategies, each subdivided in different options.

- based on the original variables
  - The original values
  - The original values rescaled as frequencies
  - An optimal scaling
- based on factor scores
  - Factors scores with same length
  - Factor scores according eigenvalues

We propose to follow 2 lines

- The original values rescaled as frequencies
- Factor scores according eigenvalues

Original values rescaled as frequencies

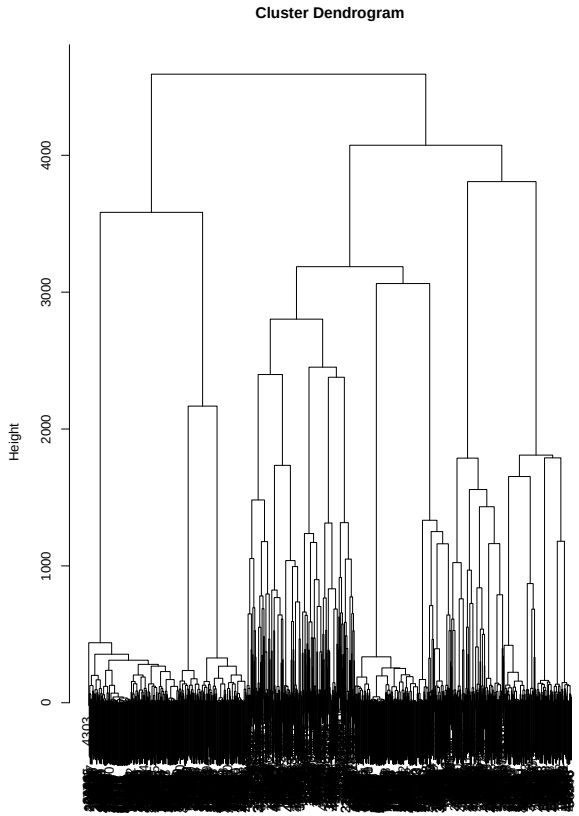
In this approach, the hypothesis is that frequency is the first criteria to take into account. That means that a daily activity like TV or internet will "count" more than a less frequent activity like cultural events for example.
The steps of the analysis will be

- A non hierarchical approach defining 20 (for example) groups
- A hierarchical analysis starting from these 20 groups
- Decision about the number of clusters to keep
- Reallocation using the previous configuration
- Interpretation of the final results

Hierarchy is not really useful

```
SCRs<-SCR[sample(nrow(SCR),1000),]
hw<-hclust(d,method="ward.D2")
```

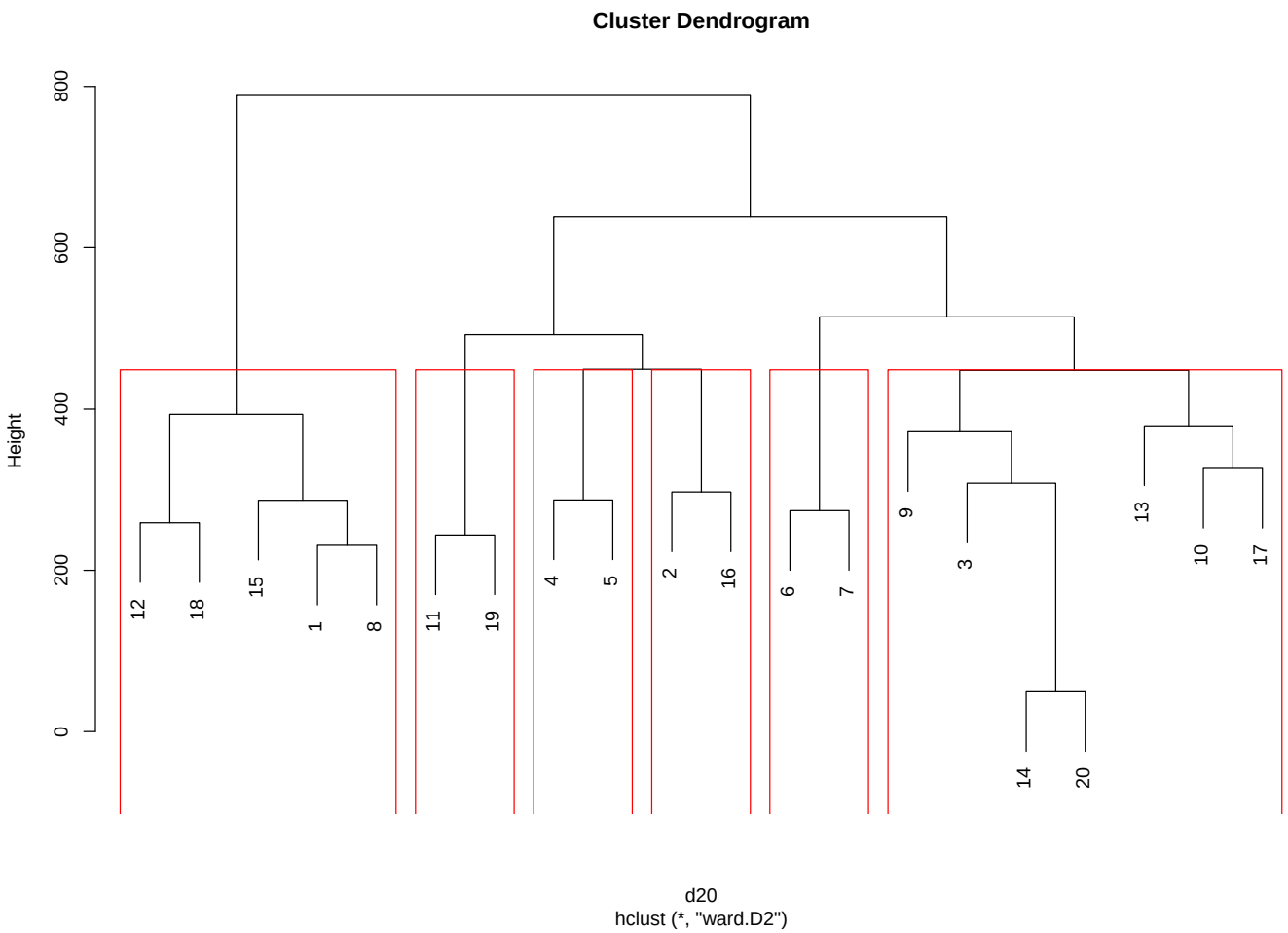
```
d<-dist(SCRs,method="euclidean")
plot(hw)
```



R code for analysis

```
#kmeans with 20 clusters
nh20<-kmeans(SCR,20)
#Distance to use for a hierarchical analysis
d20<-dist(nh20$centers)
hw20<-hclust(d20,method="ward.D2")
plot(hw20)
#Cut the tree and reallocation
group6<-cutree(hw20,k=6)
rect.hclust(hw20,k=6,border="red")
nh20$c2<-0
for (i in 1:20) nh20$c2[nh15$cluster==i]<-group6[i]
hw6c<-aggregate(SCR,by=list(nh20$c2),FUN=mean)
nh6r<-kmeans(SCR,centers=hw6c[2:14])
```

Hierarchical tree

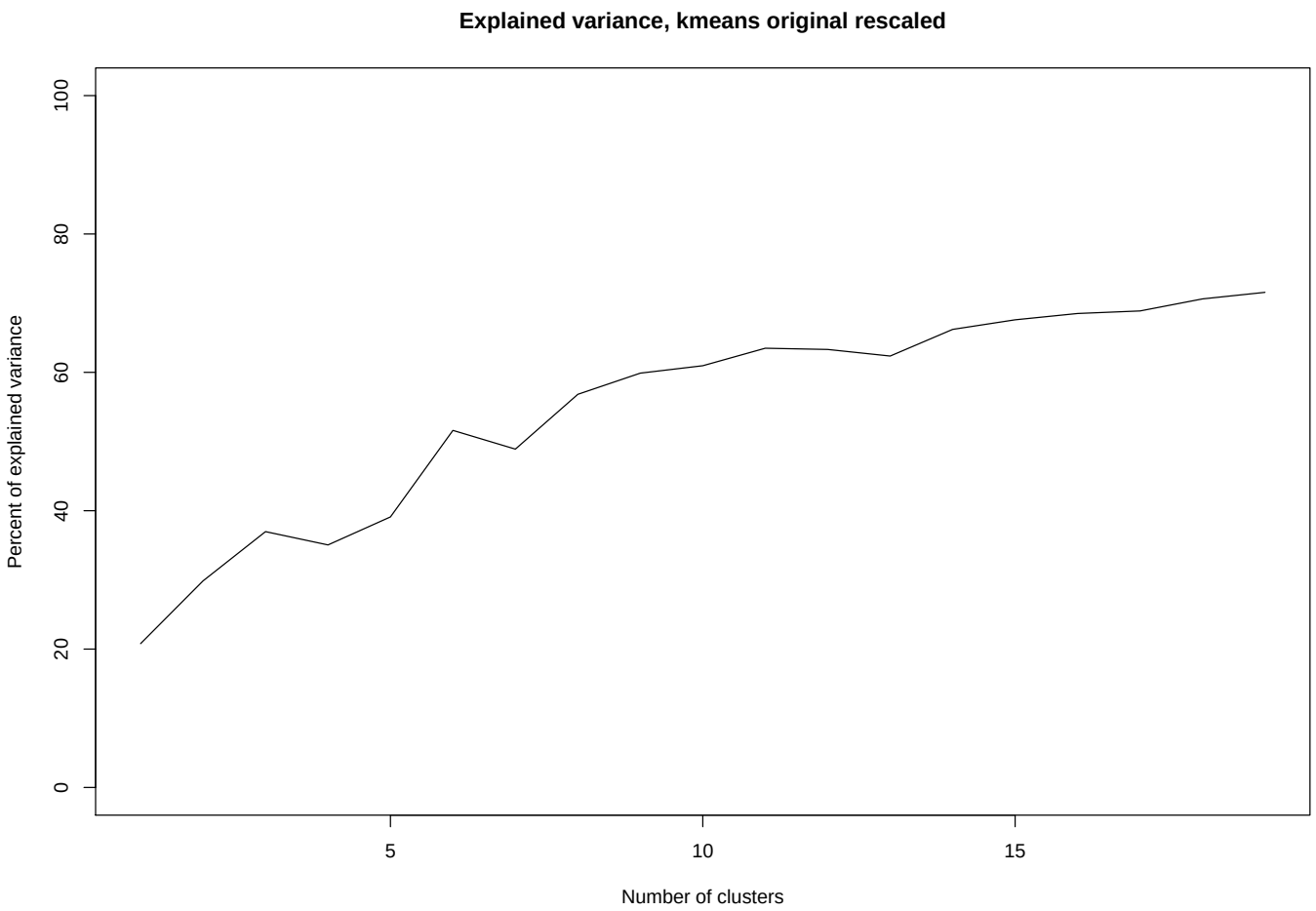


How to choose the number of clusters?

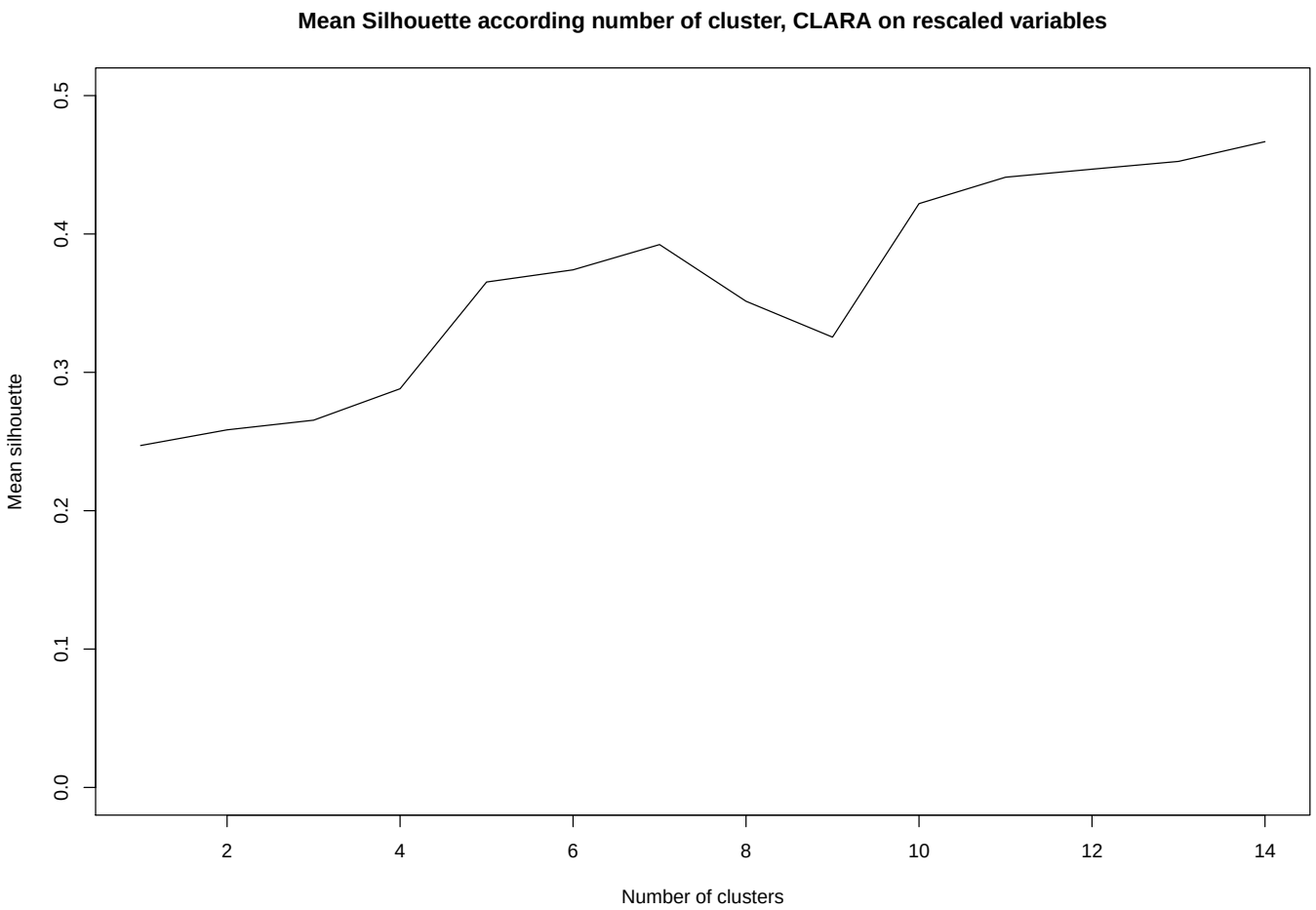
- Hierarchical tree: OK but not always easy to read
- Plot of Explained variance
- Silhouette value in the context of CLARA
- Time for interpretation
- Theory...

The first three can be computed automatically in the "factoextra" package but not difficult to implement with some lines of code in R
More discussion on the latter two...

Variance explained in kmeans



Silhouette variation in the context of CLARA



How to interpret results?

	1	2	3	4	5	6
TV						
Movies						
Shopping						
Books				++++		
Concerts						
Relatives		++++++++				
Friends		+				+++++
Cards						
Music	+		–			
Sports						
Match						
Handicraft						
Internet	-		-		+++	

Computation of factor scores

We have already mentioned the problem of scaling. A solution is to use factor scores: as PCA is based on correlation, all the variables have the same weight, independently of scale or variance.
In the case we have used all the information, meaning 13 factor scores but it could be argued to limit the analysis on the most important dimensions, the last ones being considered as "noise".
In any case, it is important to weight the factor scores according their variance by multiplying them by the square root of their eigenvalues.
In the case, according the ordinal characteristics of the variables, we have chosen a non-linear principal component analysis.

Steps for such an analysis

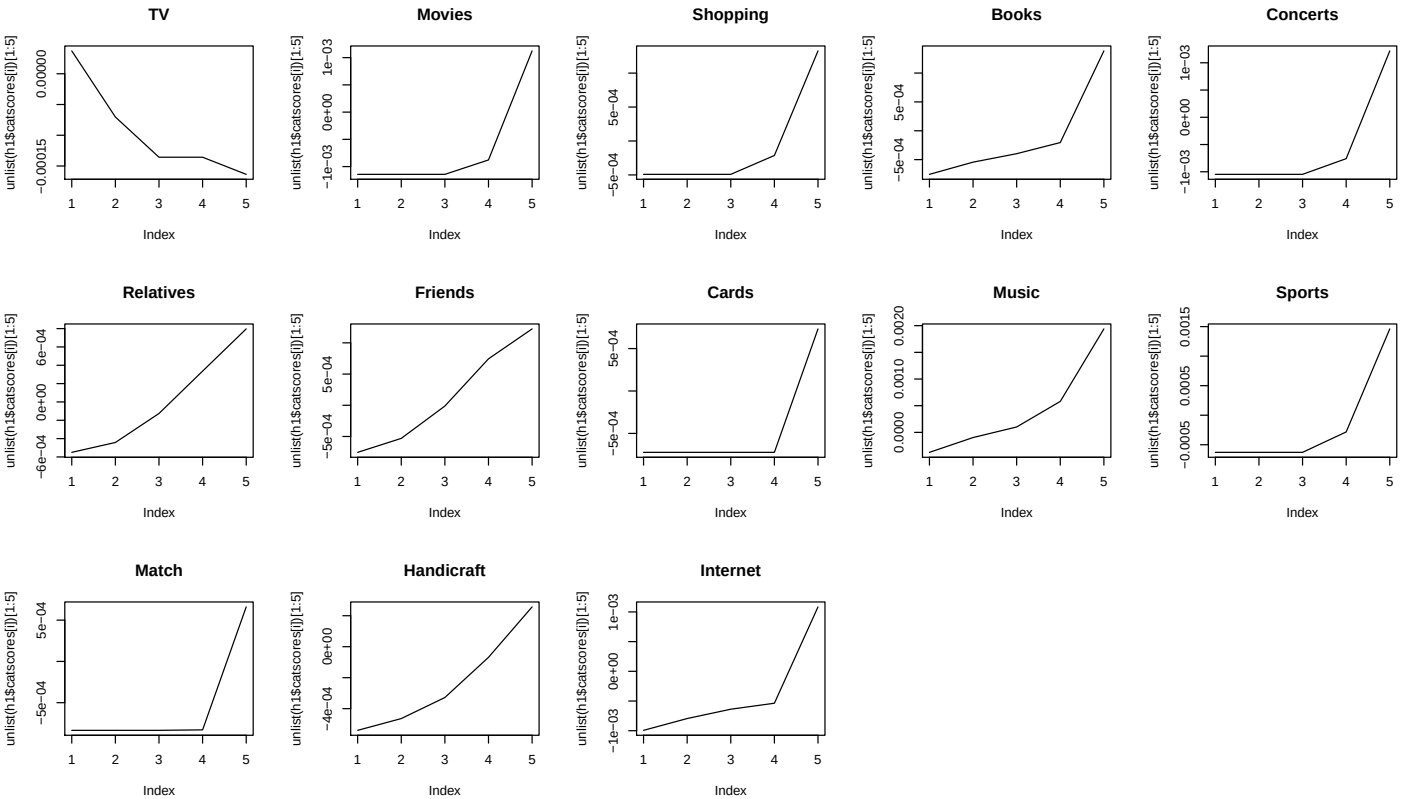
NLPCA using the homals routine developed by GIFI and save the rescaled variables from the scoremat object

```
h1<-homals(OCR,ndim=4,rank=1,level="ordinal")
reco<-as.data.frame(h1$scoremat[,,1])
```

Possible to do a classical PCA on the rescaled variables and weight the scores according eigenvalues

```
library(psych)
f1<-principal(reco,nfactors=13,rotate="none",scores = TRUE)
ei<-eigen(cor(reco))$values
ma<-as.data.frame(f1$scores)
for(i in 1:13){   ma[,i]<-ma[,i]*sqrt(ei[i]) }
```

Real example on scores



Profile of the cluster

	1	2	3	4	5	6
TV						
Movies	+	+	-	-	+	+
Shopping		+			-	+++
Books	+++	+			-	-
Concerts	+	+	-	-		
Relatives						
Friends		+				
Cards			++	-		
Music		++++++				
Sports	+	+		-		
Match		+				
Handicraft						
Internet	+	+	-	-	+	+
N	5500	3013	6502	13093	5519	3456

35 % explained variance

Interpretation

We can expect different dimensions as important for the classification obtained.

- A difference according (relative) differences: more frequent practices against less frequent ones: this is the case between cluster 1 and 2 against the 4.
- A difference between highbrow culture and other forms of leisure. This could be the case for the cluster 1, higher for books, concerts and movies than the other groups in general, and the cluster 3 in particular where games of cards are relatively more frequent.
- A difference between young and urban activities, also associated with a higher internet use. This is the case of cluster 6.

Conclusions

In summary, the following points seems important for me when using this kind of technique:

- Cluster analysis is a powerful tool for data analysis even if any researcher must be careful
- The stability of results by using different techniques is also a quality criterion
- The question of scales and measurement is very important to consider explicitly
- The use of dimensional analysis as complementary tool could be a general recommendation

Conclusions 2

- A classification is easy to communicate. At the same time it is important to have a theoretical ground for interpretation
- A classification is a categorical variable but the categories can be mapped in a dimension structure
- In summary, a very useful tool but only a part in the toolbox
- And a careful balance to consider between technical considerations and social science knowledge

To go further

As mentioned, the book of Brian Everitt [Everitt(1974)] is a very good introduction. From decades to decades, it has been completed until the fifth edition in 2011.
In the seventies and later, quite a number of interesting books were published like [Anderberg(1973)], [Hartigan(1975)] or [Romesburg(1984)] but also in French [Lebart *et al.*(1977)Lebart, Morineau and Tabard]. Later the work of Kaufman and Rousseeuw represented an important update on use of cluster analysis [Kaufman and Rousseeuw(1990)]. More recently, in particular in the French line, we can mention [Kassambara(2017)].
For an other approach on multidimensional analysis and alternative scaling see [Gifi(1990)] and for an application of alternative scaling in social sciences [Joye *et al.*(2019)Joye, Birkelund and Lemel].

References


Anderberg M. R. (1973) *Cluster Analysis for Applications*, Academic Press.


Everitt B. (1974) *Cluster Analysis*, Heinemann.


Gifi A. (1990) *Nonlinear Multivariate Analysis*, Wiley.


Hartigan J. A. (1975) *Clustering Algorithms*, Wiley.


Joye D., Birkelund G. E. and Lemel Y. (2019) *Empirical investigation of Social Spaces*, chap. Travelling with Albert Gifi: Nominal, Ordinal and Interval Approaches in Comparative Studies of Social and Cultural Spaces, pp. 393–410, Springer.


Kassambara A. (2017) Practical guide to cluster analysis in r: Unsupervised machine learning (multivariate analysis book 1).


Kaufman L. and Rousseeuw P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley.


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


Romesburg H. C. (1984) *Cluster Analysis for Reaserchers*, Lifetime Learning Publications.