Clustering Analysis

Authors: Cedric Bhihe <cedric.bhihe@gmail.com>
Santi Calvo <s.calvo93@gmail.com>

Date: 2018.04.23

1. Prepare the imputed data

As in a previous report, we read the Russet [1] data set and conduct a MICE[i] imputation. The initial data set is available as `Data/russet_imputed-values_mice-pmm.txt`. We loaded the data in the data-frame object, X, by means of our main R script available as `Lab4/script_clustering.R`.

2. PCA of continuous variables with Factominer; consider Cuba an outlier

We proceed to carry out the PCA with FactoMineR on the raw data, with scaling and:
- the "Cuba" individual as ind.sup
- the "demo" variable as quali.sup

We then check the result obtained with FactoMineR with the PCA analysis carried out manually. To do so, we apply a uniform ponderation on all observations but "cuba" which receives ponderation zero (`Data/20180319-131250_arbitrary-wparam.csv`). This is equivalent to considering Cuba an outlier. We suppressed the variable "demo" from the observation matrix X and standardized the resulting n x p matrix. We diagonalized the correlation matrix $X^T N X$, where N is the n x n diagonal ponderation matrix with zero weight for "Cuba". Result were strictly identical to those previously reported in Lab reports #2 and 3.

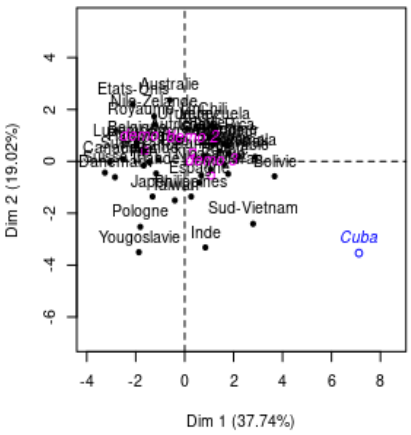
Our implementation yields identical eigenvalues as in Lab·2.

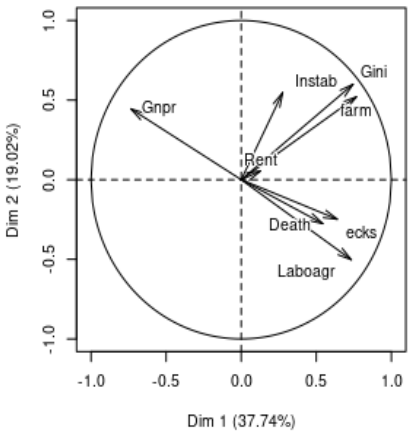
PC	Eigen-value	% of variance	Cumulative % of variance
PC 1	3.02	37.74	37.74
PC 2	1.52	19.02	56.75
PC 3	1.12	13.96	70.71
PC 4	0.91	11.43	82.14
PC 5	0.62	7.76	89.90
PC 6	0.57	7.17	97.07
PC 7	0.18	2.29	99.36
PC 8	0.05	0.64	100.00

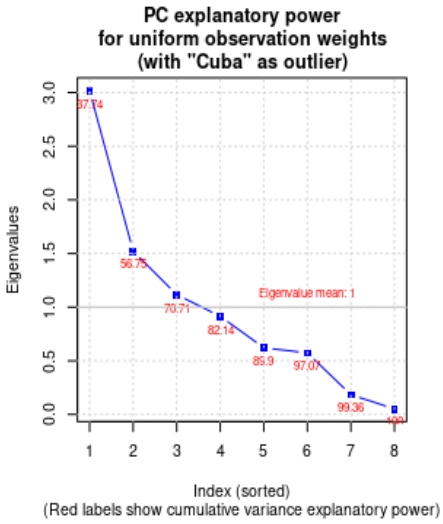
Table 1*: Eigenvalues and cumulative percentage of variance representation along principal directions (rounded within 0.01). The four significant dimensions correspond to the green shaded rows (nd=4).*

3. Select the nd significant dimensions

We choose all eigenvalues such that their their cumulated variance representativeness is >= 80% (in shaded green in Table 1). The number of significant dimensions is 4.







4. Perform an HC with the significant factors, decide the number of final classses to obtain and consolidate the clustering

First we perform a probabilistic clustering analysis using ten (10) k-means replications, while the sweeped number of clusters goes from 2 to 10. For each replicated experiment, we calculate two criteria to zero down on the optimal number of clusters:

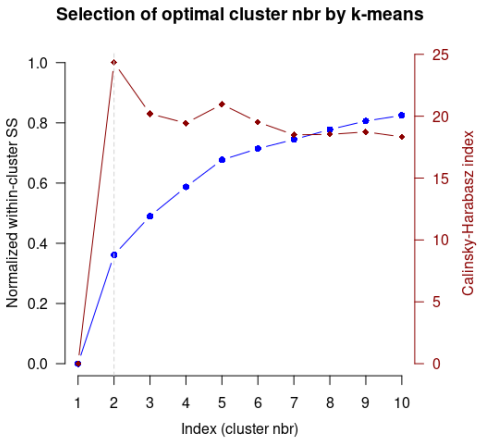
---

i MICE: Multivariate Imputation by Chained Equations

(1) within-cluster inertia's sum of squares over total inertia's sum of squares, referred to as the "normalized within-cluster SS criterion"

(2) Calinsky – Harabasz index

The resulting plot to the right suggests that the optimal cluster number (signaled by a vertical grey line) is 2.

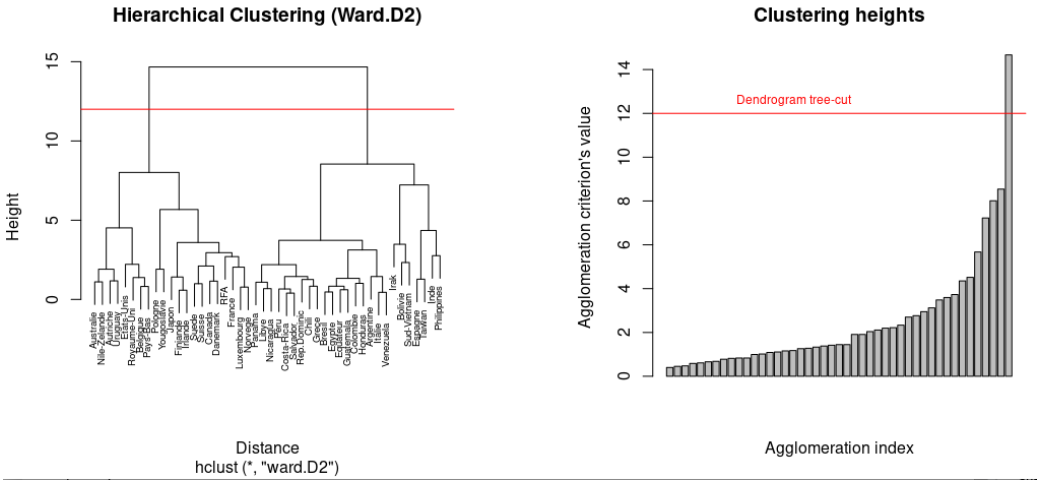


Next we proceed with a hierarchical clustering (HC) approach, followed by a k-means consolitdation. Input for the clustering algorithm is the Euclidian distance matrix based on ψ, itself defined in earlier reports as the projections of obervations on the principal directions. We relied on the ward.D2 distance measure. It takes into account the inertia associated with each cluster's centroid, i.e. each cluster's centroid is weighted by the same-cluster observations surrounding it.

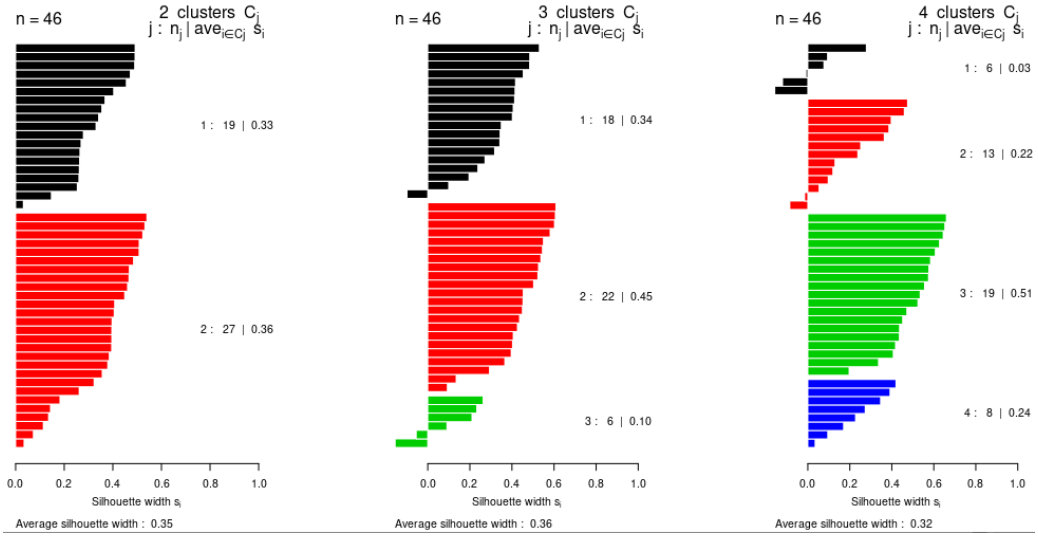
We observe from the dendrogram and heights-of-joins plots that the number of cluster could be 2, with centroids G1 and G2 shown in Table 2

	G1	G2
PC1	1.39	-1.65
PC2	-0.12	0.14
PC3	-0.02	0.02
PC4	-0.14	0.17

***Table 2**: Centroids coordinates for the HC.*

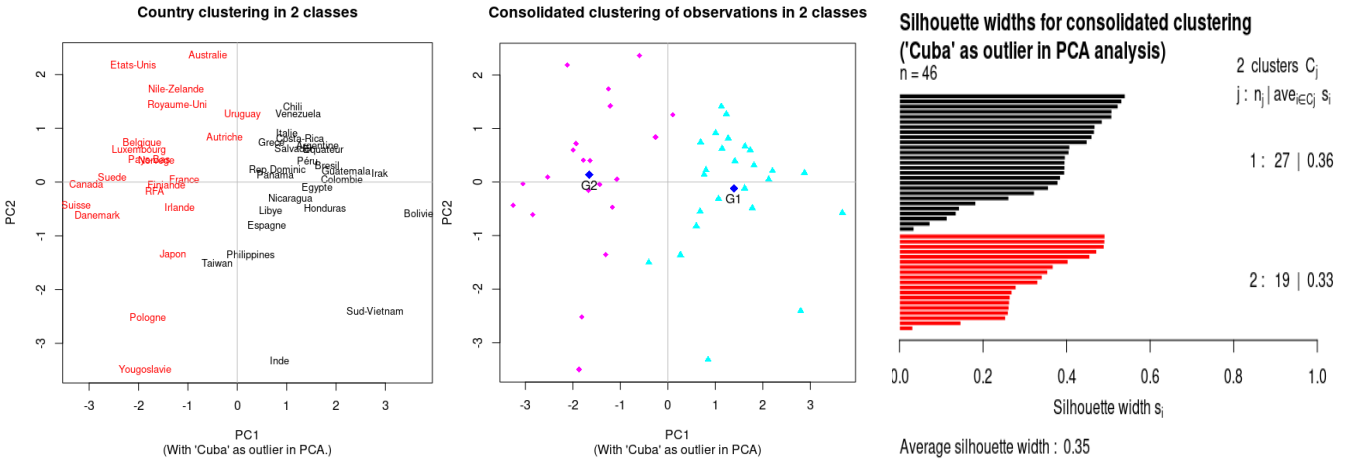


The silhouette method applied to 2, 3 and 4 clusters seems to confirm that, with positive S-widths for 2 clusters and some negative ones for 3 and 4, suggesting that partition numbers higher than 2 yields errroneous categorization in certain cases.



From this first clustering analysis stage, we obtain a *quality index* ***Ib=35.59*** (within 0.01).

We visualize this first partition result in the first factorial plane per the next figure (below left). The consolidation result is represented (below center), based on the k-means method, with the 2 centroids, and an *improved quality index of* ***Ib=36.16***. The silhouette widths (below right) for 2 clusters is not significantly different from the previous one.



The two detected groups seem to match on one hand countries (**Cat. 1** in turquoise) with greater instability and on the other hand (**Cat. 2** in red) countries with more political stability and wealth. Those categories and their interpretation in terms of variables are further discussed in Section 5.

5. Using catdes() to interpret clusters

We perform a categorical description based on previous cluster information (categorization for the observations).
For each category (Cat. 1 and Cat. 2) we show for each variable the test statistics based on the Student-t for each variable's significance.

V.test is the number of std-dev below or above the overall mean values for the sample data. Values of of v.test <0 mean that the category mean is smaller than the overall average and vice versa.

The null hypothesis, for continuous variables, is that the variable's mean in the given group is equal to the variable's global data mean. We reject the null hypothesis at the risk 0.05 of being wrong when the p-values <0.05. In the two Tables 3a and 3b below, we only show variables for which we reject H_0, i.e. for which the categorization is meaningful.

Cat_1	v.test	Mean in category	Overall mean	Sd in category	Overall sd	p.value
farm	4.947159	96.95185	92.82609	3.187751	6.668988	0.0000007530454
Laboagr	4.661164	55.14815	42.43478	14.567096	21.811098	0.0000031442679
Gini	4.278413	78.86667	71.19783	10.177790	14.333701	0.0000188230611
ecks	3.353320	30.03704	21.60870	20.503805	20.099187	0.0007984840288
Death	2.217522	125.07407	73.58696	228.721394	185.670065	0.0265874093464
Gnpr	-4.985790	258.74074	563.56522	159.908213	488.908040	0.0000006170909

Cat_2	v.test	Mean in category	Overall mean	Sd in category	Overall sd	p.value
Gnpr	4.985790	996.7368421	563.56522	471.901065	488.908040	0.0000006170909
Death	-2.217522	0.4210526	73.58696	1.138595	185.670065	0.0265874093464
ecks	-3.353320	9.6315789	21.60870	11.671876	20.099187	0.0007984840288
Gini	-4.278413	60.3000000	71.19783	12.160506	14.333701	0.0000188230611
Laboagr	-4.661164	24.3684211	42.43478	17.150144	21.811098	0.0000031442679
farm	-4.947159	86.9631579	92.82609	5.888455	6.668988	0.0000007530454

Table 3*: Most significant variables in each categories for (a) Cat.1 with more instability and poverty and (b) Cat. 2, countries with more stability and welath*

Table 4 (to the right) summarizeds square correlation coefficient (*Eta2*) and p-values of the F-test in a one-way ANOVA (assuming homoscedasticity) for significant continuous variable globally, i.e. for variables, whose corresponding p-value is smaller than 0.05.

In practice it is preferable to consider a smaller risk level of 0.01. In this case this would lead us to reject variable "*Death*" in addition to "*Rent*" and "*Instab*" as not significant overall.

	Eta2	P-value
Gnpr	0.5524023	0.000000003302162
farm	0.5438751	0.000000005037780
Laboagr	0.4828099	0.000000084441780
Gini	0.4067737	0.000001866647754
ecks	0.2498834	0.000404416982041
Death	0.1092757	0.024844049542374

***Table 4**: Summary of significant variables' correlation and p-values globally. The "Death" variable's is shaded because we reject it as non significant based on a risk level of 0.01.*

6. Assign Cuba to one of the defined clusters

We assign "*Cuba*" to Cat. 1 as its coordinates in the first factorial plane are: (7.121731, 3.536060).
The corresponding point is located far to the right and in the lower quadrant of the PC1-2 factorial plane representation of the points.

References:

[1] B. M. Russet, "Inequality and Instability", World politics, 21 (1964), 442-454.

[2] H. Wold, "Estimation of principal components and related models by iterative least squares," in Multivariate Analysis (Ed. P.R. Krishnaiah), Academic Press, NY (1966), 391-420.