Multivariate Analysis

Principal Components Analysis Report

Authors: Cedric Bhihe <cedric.bhihe@gmail.com> Last edit: 2018.03.21

1. Prepare the data.

We read the Russet[i] data set and conduct a MICE[ii] imputation so as to rid our data of any missing values, following the the same processing steps as in Lab1 (predictive mean matching (pmm) imputation). That initial data set is available as *Data/*`russet_imputed-values_mice-pmm.txt`. It is loaded as the data-frame object, *X*, in R by means of our main R script available as `Lab2/script-pca.R`.

2. Write a function for PCA analysis, allowing for both the Euclidean metric and the normalized Euclidean metric.

Our function `pcaF(X,wflag,wparam,...)` accepts the following inputs:

- *X,* the imputed data frame for observations, later either centered or standardized.
- *wflag,* taking values in `c("uniform","arbitrary","random")` in such a way that observations can be variously ponderated (viz): uniformly, arbitrarily (e.g. by setting an individual's ponderation to 0 while all others remain at 1), randomly (which is a "toy option" with no practical statistical *a priori* interest, other than to see "*what happens if …!*").
- *wparam*, an n-dimensional vector containing the individuals' ponderation when *wflag* is `"arbitrary"`.

In practice any arbitrary individuals' ponderation vectors can be produced (and saved to disk for later retrieval) by means of a short secondary script available as `Lab2/wparam-arbitrary.R`.

Results for uniform individuals' ponderation:

Produce weight matrix, N: `W ← rep(1,nrow(X)); N ← diag(W/sum(W),nrow(X),nrow(X))`

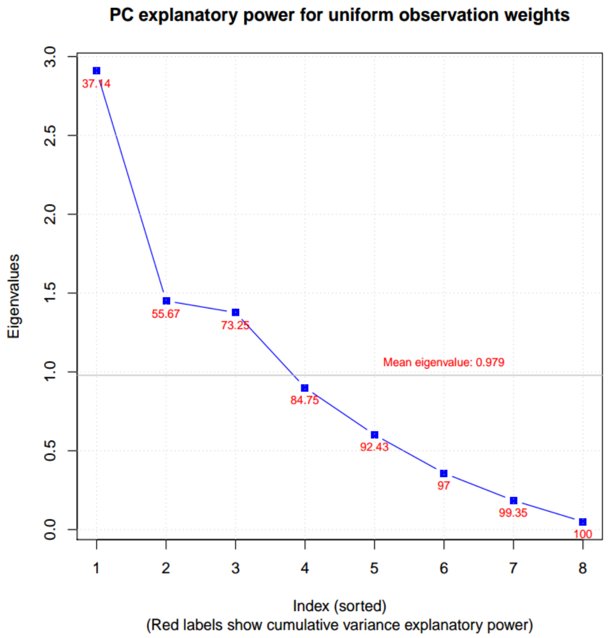
Calculate the centroid: `(centroid <- apply(N %*% as.matrix(X), 2, sum))`

```
Gini     farm     Rent     Gnpr      Laboagr  Instab   ecks     Death
71.36809 92.93191 20.90213 559.25532 42.42553 12.40851 23.27660 133.72340
```

The centroid is used to center and to standardize observations. We calculated the eigenvalues from both the covariance (centered obs.) and the correlation (standardized obs.) matrices, both with rank 8. We found that in the former case (cov. matrix for centered data), not normalizing data produces an analytical artefact in the form of two prevailing eigenvalues, to the detriment of other principal components (PCs). As a general rule, we recall that normalizing observations is preferable for continuous variables when they are expressed in different units. **In this work, we work exclusively with standardized (i.e. normalized) data**, and with no correction for sample size. The latter introduces numerical uncertainty on the second decimal of eigenvalues.

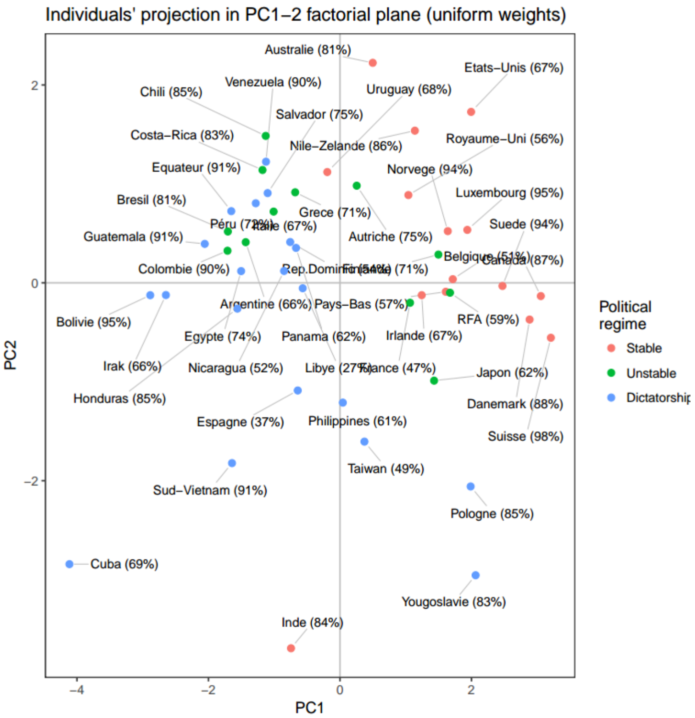
We retain 3 significant PCs with eigenvalues 2.91, 1.45 and 1.38 (using Kaiser rule with a threshold fixed at eigenvalue mean 0.98). After rounding, they account respectively for ~37.1%, ~18.5% and ~17.6% of variability information in the p=8 dimensional variables' space. The screeplot to the right includes the cumulative explanatory power of the 3 significant PCs as red labels. The total retained information is ~73.3%.

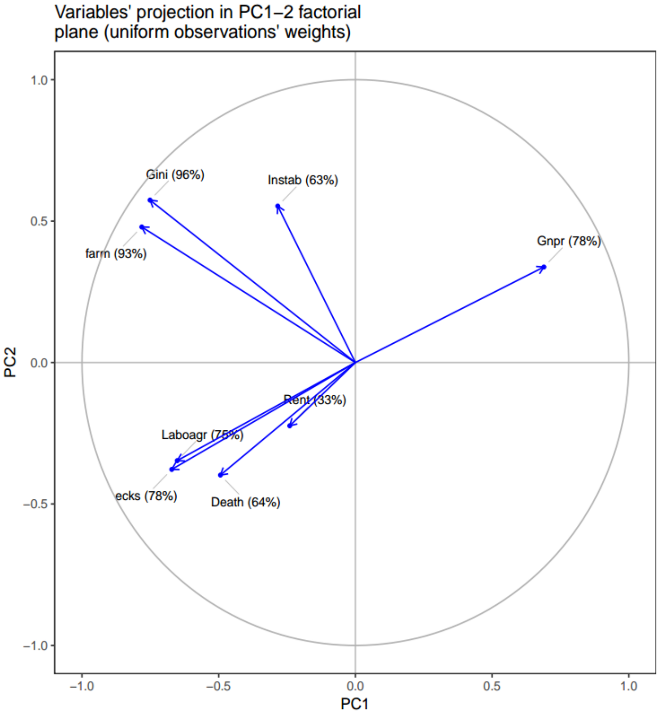
Further (below) we present both the observations' projection and the variables' projection in their respective PC1-2 factorial spaces (hyperspaces in $\Re^8$ and $\Re^{47}$ respectively). Projections in PC2-3 and PC1-3 factorial spaces are also provided but not included in this report. All observations' projections coordinates are kept in `Lab2/Report/[datestamp]_psi_unif.csv` for later retrieval. In both plots, labels indicate the projection's fractional representativeness (in % respectively for individuals and for variables) in their PC1-2 factorial planes. For variables those percentages scale as expected with arrow length.



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

[ii] MICE: Multivariate Imputation by Chained Equations





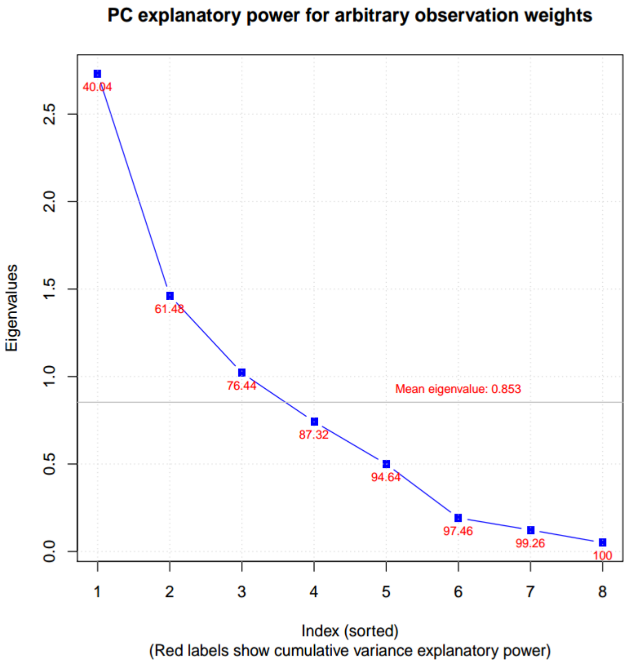
	PC1	PC2	PC3
Gini	-0.7512	0.5742	-0.0419
farm	-0.7814	0.4792	-0.0357
Rent	-0.2403	-0.2236	-0.6445
Gnpr	0.6894	0.3377	-0.4964
Laboagr	-0.6518	-0.3470	0.5642
Instab	-0.2840	0.5532	-0.1175
ecks	-0.6717	-0.3779	-0.2885
Death	-0.4937	-0.3978	-0.5440

Examining the projections of all 8 variables in the 3 significant PC dimensions, we see that along PC1, the two highest and are ”farm” (-) and “Gini” (-) in salmon and the two lowest correlations are “Rent” (-) and “Instab” (-) in green, all negatively correlated. Others (PC2 and PC3) follow different patterns and are identically color-coded. Data is available at: `Lab2/Report/20180320-085019_phi_direct_unif.csv`.

3. Results for Cuba observation’s weight set to zero

The corresponding *arbitrary* weight vector is available at: `Data/20180319-131250_arbitrary-wparam.csv`. The newly calculated centroid is slightly displaced to:

```
    Gini     farm     Rent     Gnpr      Laboagr  Instab   ecks     Death
    71.19783 92.82609 20.18696 563.56522 42.43478 12.38261 21.60870 73.58696
```

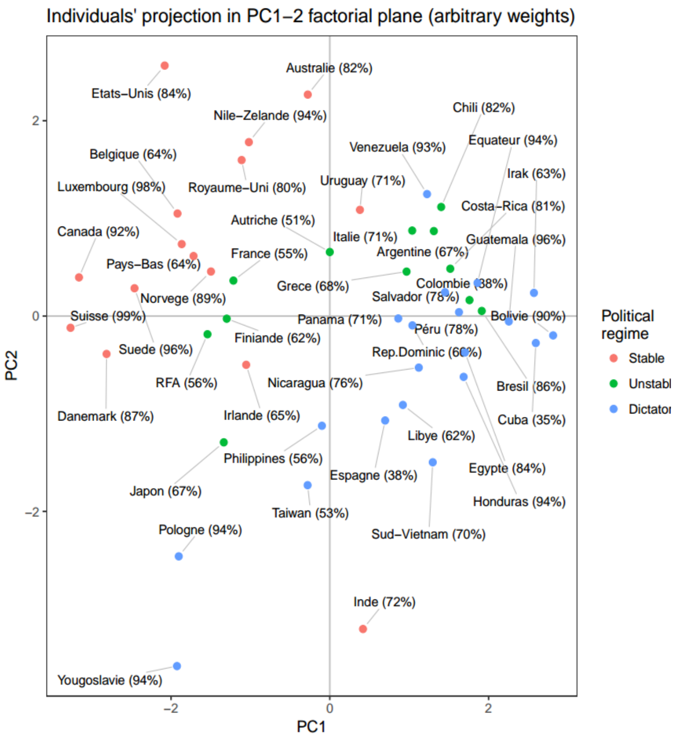


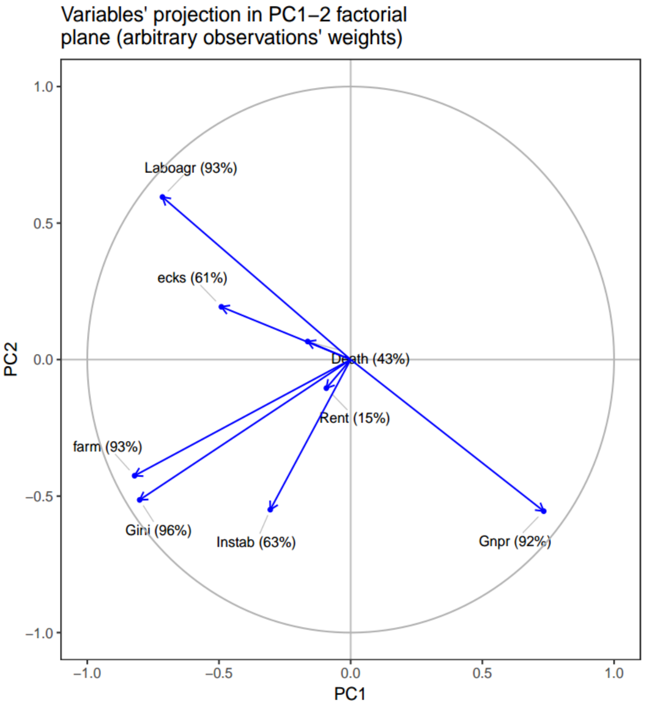
The centroid displacement is primarily due to the diminished effect of the “Death” variable. It large value in the Cuba observation may explain its tentative classification as outlier.

The new correlation matrix rank is 7. Following Kaiser rule, we retain 3 significant PCs, with a slightly increased cumulated variance explanatory power of 76.4% (red labels) and eigenvalues: 2.73, 1.46, 1.02. They correspond to the variability representativeness of PC1 40.0% , PC2 21.4% and PC3 15.0%. An increase in cumulated explanatory power is always welcome, but may not justify in the end ignoring an observation.

	PC1	PC2	PC3
Gini	-0.8012	-0.5139	0.0498
farm	-0.8204	-0.4253	0.1444
Rent	-0.0924	-0.1046	0.8528
Gnpr	0.7329	-0.5552	0.0358
Laboagr	-0.7146	0.5956	-0.1515
Instab	-0.3052	-0.5497	-0.4876
ecks	-0.4909	0.1933	0.0811
Death	-0.1632	0.0662	0.0389

Changes are also visible in the correlations of variables with the newly retained PCs. For PC1 “farm” (-) and “Gini” (-) remain the highest two correlated variables. For PC2, “Gini” (-) is replaced by“Laboagr” (+) and “Instab” (-) remains. For PC3, “Rent” (+) remains and “Instab” (-) replaces “Laboarg” (+) . This is further mirrored by the two next graphs.





Variable projection data is available at: `Lab2/Report/20180320-085020_phi_direct_arbi.csv`.

As before plots for observation and variable projections are also available in the two factorial planes PC1-2 and PC1-3 in the folder `Lab2/Report`.

4. Sensitivity of the PCA with respect to considering Cuba as an outlier

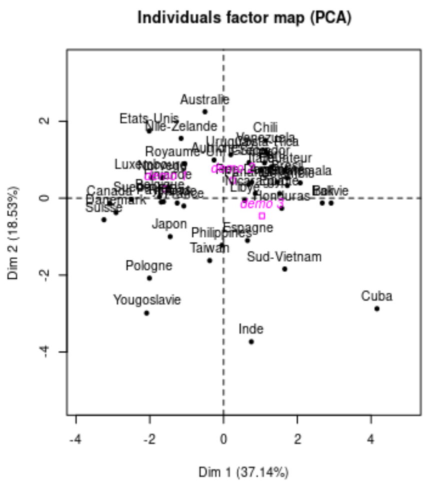
We perform the correlation on the matrices of individual projections reduced to their 3 first columns (3 significant PCs) for arbitrary weights (`psi_arbi` where Cuba weight → 0) and uniform weights (`psi_unif`).

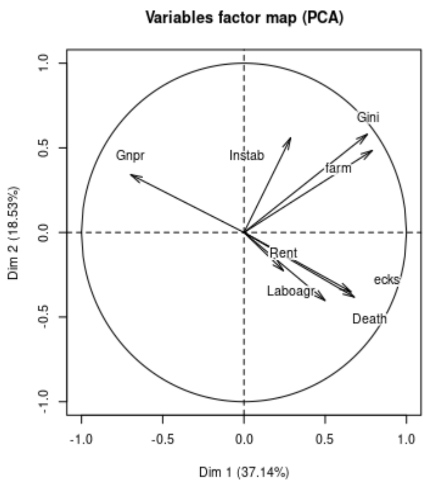
```
diag(cor(psi_arbi[,1:3],psi_unif[,1:3]))
    PC1     PC2     PC3
-0.9857  0.8792 -0.6133
```

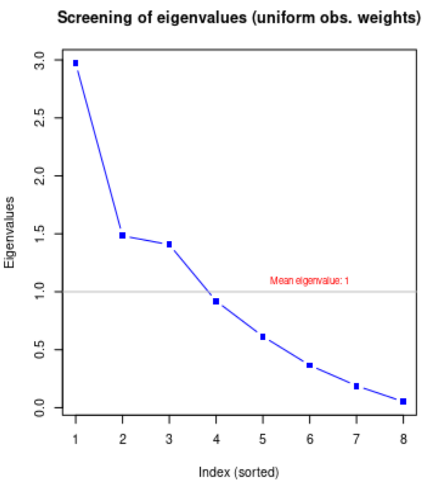
We detect a strong correlation (negative) in the PC1 direction and a moderate to low correlation for PC2 and PC3. For that reason we chose to retain Cuba as active observation in the rest of this work with FactoMineR.

5. Study with FactoMineR

We plots individuals and variable projections in factorial plane PC1-2, along with eigenvalues' screeplot and 3 significant PCs according to Kaiser's rule. The cumulative explanatory power of the significant PCs is 73.3%, with demo as illustrative.







As before the variable group {Death, Laboarg, ecks} on one hand and Gnpr on the other hand are strongly anticorrelated, while Gini, farm and to a lesser extent Instab form a separate correlation group.

To exhibit countries best represented in PC1-2, we compute the individuals' projected norm to unprojected norm (Frobenius) ratios. Results are: ***Switzerland 97.8%, Bolivia 97.3%, and Luxembourg 96.9%***.

The three worst represented in PC1-2 are: ***Taiwan 50.0%, Spain 37.5% and Libya 28.6%.***

The 3 countries most influencing PC1 are: ***Cuba 4.16, Switzerland -3.24, Canada -3.09***

```
psiFMR <- pcaX$ind$coord
for (ii in 1:3) { cat(rownames(psiFMR)[order(abs(psiFMR[,1]),decreasing=T)]
[ii],round(psiFMR[order(abs(psiFMR[,1]), decreasing=T)][ii],2),"\n") }
```

Similarly, the 3 countries most influencing PC2 are: India -3.73, Yougoslavia -2.99, Cuba -2.87

The ranking of variable representation in PC1-2 shows that Gini is best and Rent is worst.

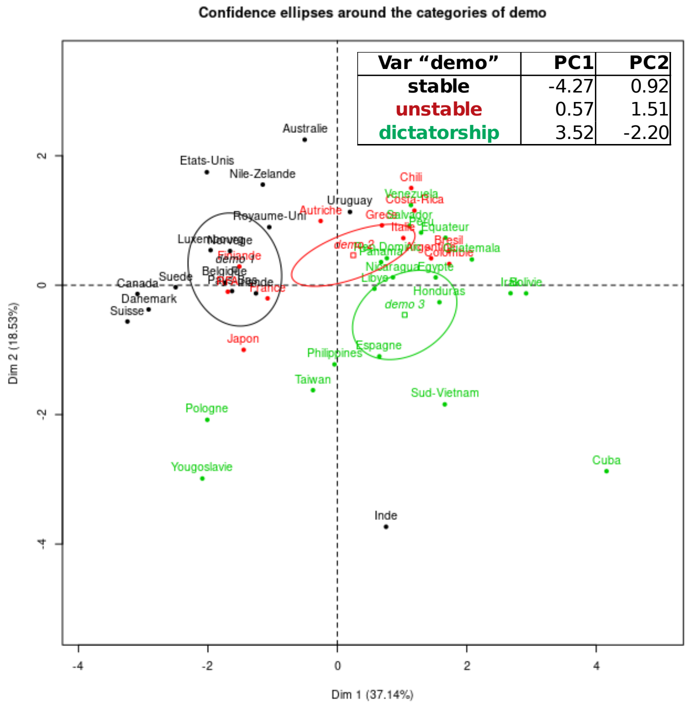
Gini	farm	ecks	Gnpr	Laboagr	Death	Instab	Rent
96.86%	93.88%	87.23%	81.89%	78.46%	69.53%	62.9%	33.21%

The 3 variables most influencing the formation of PC1 are: ***farm 0.79, Gini 0.76, Gnpr -0.70***

```
phiFMR <- pcaX$var$cor
for (ii in 1:3) { cat(rownames(phiFMR)[order(abs(phiFMR[,1]),decreasing=T)]
[ii],round(phiFMR[order(abs(phiFMR[,1]), decreasing=T)][ii],2),"\n") }
```

Similarly, the 3 variables most influencing the formation of PC2 are: ***Gini 0.58, Instab 0.56, farm 0.48***

To ascertain that declared modalities ("demo") are significant along the first 2 principal directions, examine the value of the object returned by PCA in FactoMineR. $quali.sup contains a list of matrices with results pertaining to the supplementary categorical variables (i.e. to their projected coordinates along each axis of interest). Within that list, the value-test (v.test) is of particular interest[i]. It is the result of a test of the hypothesis that the mean of the modality group's individuals' projected coordinates' distributions is equal to the overall mean.



Var "demo"	PC1	PC2
stable	-4.27	0.92
unstable	0.57	1.51
dictatorship	3.52	-2.20

Lower p-values coincide with higher v.test values and denote a more significant modality representation along a given axis, assuming that distribution of the categories' individuals' projected coordinates follow a normal distribution. The v.test value is expressed in unit of standard deviation. |v.test| > 1.96 means p < 0.05. It is negative if the mean of the category's projected coordinates distribution along one axis is lower than the corresponding global mean, and vice versa.

Here *stable* and *dictatorship* are significantly represented on the PC1 axis, while only *dictatorhip* is, on axis PC2.

We show graphically (left) how potential clusters corresponding to the three declared modalities of the suplementary (categorical) variable "demo" are separable neither on the first axis (PC1) nor on the second (PC2). Their barycenters are represented by square symbols. When applying a hierarchical clustering algorithm (HCPC) with a minimum of 3 cluster, 2 non trivial clusters are discovered, while the third solely consists of the Cuba individual. A preliminary conclusion is that hierarchical clustering results do not conform with the declared modality (demo variable), raising a question as to its significance as a strong discriminant.

[i] A. Morineau, "Note sur la caractérisation statistique d'une classe et les valeurs-tests.," Bulletin Technique du Centre de Statistique et d'Informatique Appliquées, Vol 2, n° 1-2, p 20-27 (1984), available at http://www.deenov.com (in French).