Beyond Principal Components Analysis

MIRI / MVA - Lab #3Report

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

Delivery: before 2018.04.02 – 23:55

1. Prepare the data.

As in our previous Lab·#2 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 `Lab3/lab3-script_mva.R`. Our first step was to standardize our n x p observations matrix, X. We already know from previous work that its number of significant dimensions is nd = 3. All this laboratory was conducted retaining Cuba in our data based on correlations reported in Lab #2 between individual projections on the 3 first PCs for the analyses carried out with and without Cuba.

2. PC determination with NIPALS[ii]

Given the standardized representation, $X \leftarrow X/S = N^{-1/2}\,U\,V^T$, of a p-dimensional set of n observations X with uniform observations' weight square diagonal matrix N, the NIPALS algorithm [2] extracts one principal component (PC) at a time, from the iterative regression of X on the scores, $\psi = X V = U$. The pseudo-code (see frame to the right), as we implemented it in R applies to data without missing values, i.e. imputed data.

Our implementation yields identical eigenvalues as in Lab·2. We only report eignvalues for the nd=3 most significant dimensions: 2.91 (37.1%), 1.45 (55.7%), 1.38 (73.2%) with cumulative variance representativeness shown in parentheses.

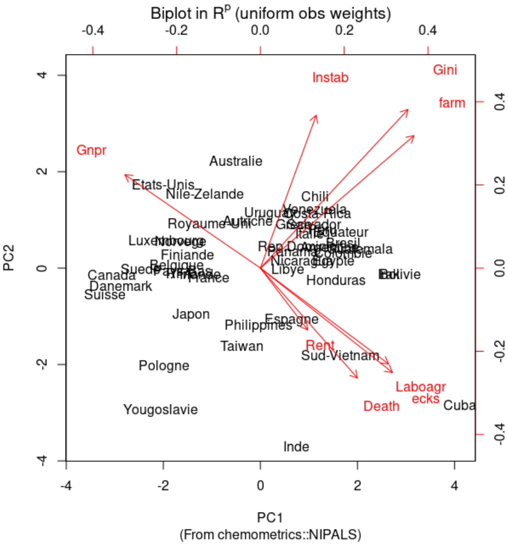
NIPALS algorithm's pseudocode

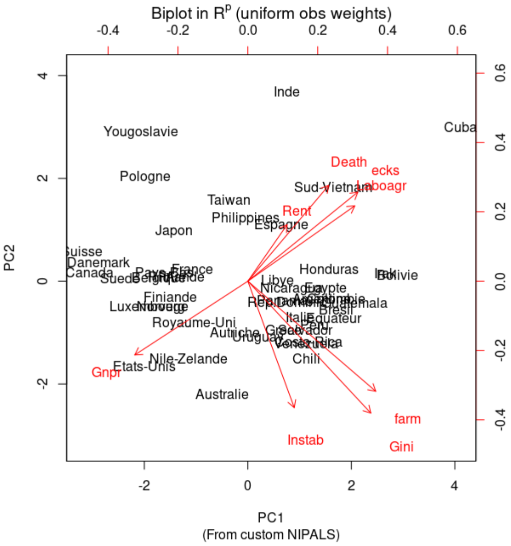
```
Loop over d, with d in 1:nd, such that nd ≤ min (n,p)
- choose eigenvalues' convergence threshold, $\varepsilon$ = 1e-5
At i =0:
    - choose $u_{(0)}$ as the column of X (standardized),
      whose sum of squared components is greatest,
    - let $\lambda_{(0)} \leftarrow 0$
    Loop over i, with i = +1:
        - compute a better loading factor, $v_{(i)}$:
          $v_{(i)} \leftarrow X^T u_{(i-1)} / (u^T_{(i-1)} u_{(i-1)})$
        - normalize the loading factor to 1:
          $v_{(i)} \leftarrow v_{(i)} / (v^T_{(i)} v_{(i)})^{1/2}$
        - re-calculate the standardized scores:
          $u_{(i)} \leftarrow X v_{(i)} / (v^T_{(i)} v_{(i)})$
        - calculate the iterated eigenvalue as scaled
          variance of the corresponding PC:
          $\lambda_{(i)} \leftarrow u^T_{(i)} u_{(i)} / n$   or   $\lambda_{(i)} \leftarrow u^T_{(i)} u_{(i)} / (n-1)$
          *(correspondence between SVD and PCA)*
    Next i until, at the k^th iteration:
          $|\lambda_{(k)} - \lambda_{(k-1)}| \leq \epsilon$
    Keep $\lambda_{(k)}$, $u_{(k)}$ and $v_{(k)}$.
    Remove the calculated PC component:
      $X \leftarrow X - u_{(k)} v^T_{(k)}$
Next d in loop.
```

3. PCA biplot in R^p

In that representation the scores of individual observations and the loadings of each variables are shown on the same PCA1 vs PCA2 chart, heretofore denoted *PCA1-2 biplot* for short. Left and bottom axes show PCs' scores while top and right axes show variable loadings.

In the present case the 2 first components represent slightly less than 56% of inertia in the observation space.





[i] MICE: Multivariate Imputation by Chained Equations

[ii] NIPALS does not refer to the Sweedish mountain [68°89'00 N, 18°30'12 E] but to the Nonlinear Iterative Partial Least Squares algorithm used to calculate eigenvalues and eigenvectors.

We note that our custom and the packaged implementations of NIPALS yield similar results. A simple change of sign along the PC2 axis is visible, the which affects neither the outcome of the scatter plot, nor its interpretation.

Meaning of the biplot:
Variables' arrows in the standardized data biplot do not only give directions of growth. They represent loadings. In R^p the biplot does not accurately reflect the correlation between the unit-scaled PCs and the original variables' projections in the PC1-2 plane. Their positions relative to one another are only grossly reminiscent of correlations between variables. Hence the cosine of their angle cannot be accurately used as a measure of their degree of correlation. Similarly their lengths only grossly reflect the standard deviation of original variables (their squared length grossly approximate their variance) scaled down from a maximum length of 1 (for a standardized observations' matrix, X).
In R^n space on the other hand, in any chosen factorial plane, arrow lengths would indicate the (percentual) degree to which the variable's projection in that factorial plane contributes to the representation of inertia/observation variability along any combination of the two principal directions defining said factorial plane.

Interpretation of R^p biplot visualization:
Bearing in mind as previously noted that angles between projected variables are not accurately represented in this biplot, we nevertheless remark that:

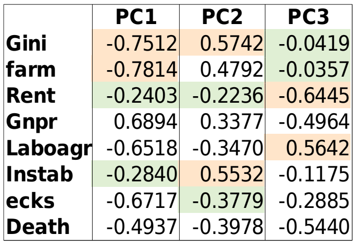
- Countries commonly perceived as stable and wealthy, i.e. with high values of Gnpr are located opposite to countries with high values of violent Deaths, larger numbers of violent conflicts (ecks) and larger percentages of their population engaged in agriculture (Laboagr).

- There seems to be an approximate anticorrelation between variables Gnpr and the group made up by Death, ecks and Laboagr. Rent is also *apparently* anticorrelated with Gnpr but its overall contribution to the representation of inertia/variability in PA1-2 *appears* too low (its arrow is *apparently* too short relative to other arrows) to be able to draw worthy conclusions in this case. Roughly orthogonal to all previously cited variables is the group of variables Instab, Gini and farm. This may suggest that a high inequality index (Gini) is *approximately* correlated with greater inequality distributions of farm land (farm) and *apparently* less so correlated with greater political instability, represented by a higher turnover of politicians in high office (Instab).

- The biplot further suggests that there may be *2 large forces acting independently, i.e. with apparently orthogonal variables' projections on the PCA1-2 factorial plane*.

- Interestingly and with the exception of Instab, none of those variables are aligned more with either PCA1 or PCA2. We showed in Lab #2 that variable Instab exhibit a modicum of preferential alignment with PC2 compared to others (as attested by the table to the right.)

- With the exception of Instab all variables seem to lie at roughly 45° from either principal axes. This suggests that no variable has a strong preferential correlation with either principal direction in the 1st factorial plane. The biplot representation as it stands could benefit from a VARIMAX operation aimed at aligning main variable groups with axes of maximum and minimum variability.



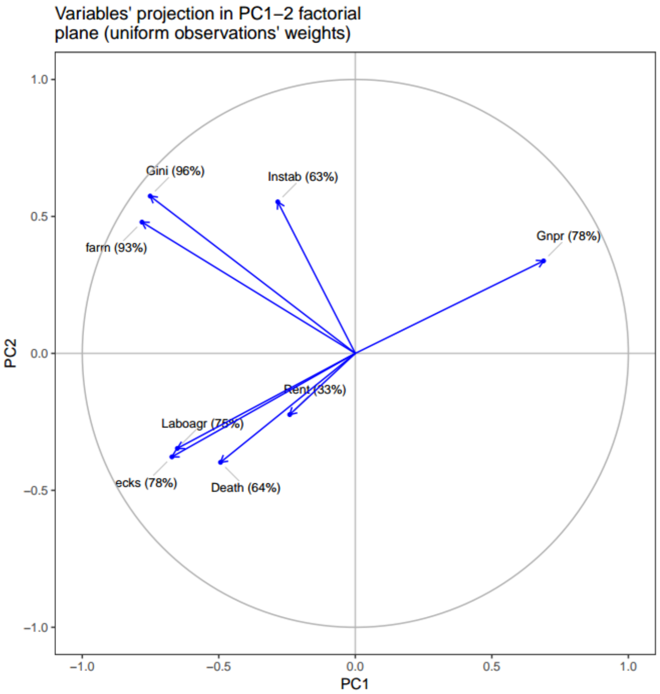
	PC1	PC2	PC3
Gini	-0.7512	0.5742	-0.0419
farm	-0.7814	0.4792	-0.0357
Rent	-0.2403	-0.2236	-0.6445
Gnpr	0.6894	0.3377	-0.4964
Laboagr	-0.6518	-0.3470	0.5642
Instab	-0.2840	0.5532	-0.1175
ecks	-0.6717	-0.3779	-0.2885
Death	-0.4937	-0.3978	-0.5440

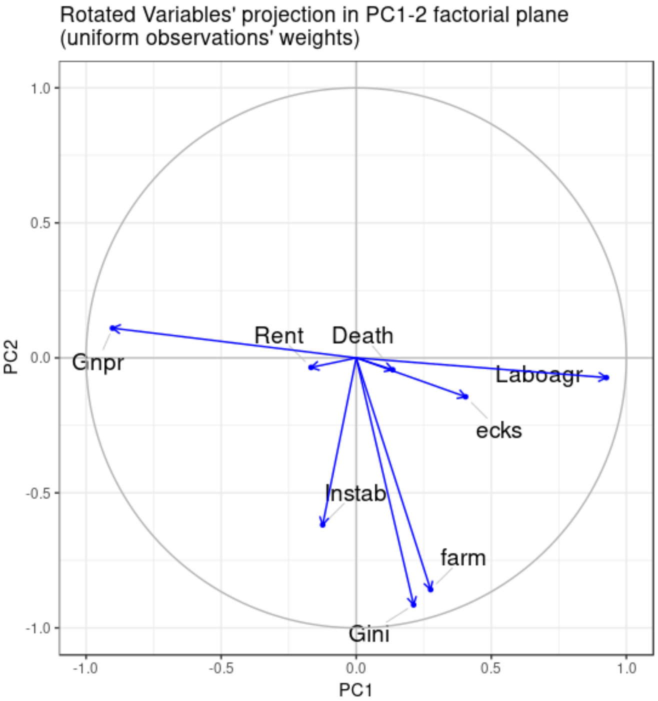
4. Plot of rotated variables with VARIMAX and comments.

We run the orthogonal rotation method, *VARIMAX*, on the "p x nd" (nd < p) matrix of projected variables, known as loadings $V \cdot \Lambda^{1/2}$ where V is the "p x nd" PCA matrix of significant eigenvectors and $\Lambda^{1/2}$ the "nd x nd" diagonal matrix, whose elements are the significant eigenvalues' square-roots. We verify that the sum of the sum of *SS loadings*[i] is equal to the sum of the eigenvalues. In our case each rotated factor (corresponding to the new rotated PC1 and PC2 basis) account for about 25% of variance each, while the third rotated factor accounts for 21.5%. The three rotated Principal axes account for a cumulated 72%, close but not exactly the cumulated explanatory power of 73% derived from our NIPALS eigenvalue computation (see Section 3).

Plots are shown below for PC1-2 variables projections before and after *VARIMAX* rotation. As could be expected for the rotated basis, we observe that Gnpr and Laboagr are almost perfectly anticorrelated. They almost coincide with the new rotated PC1 axis, consistent with the latent factor "**Wealth**". Rent and Death are preferentially correlated to the third principal (rotated axis) and thus exhibit very short projection arrows. Variables farm, Gini and Instab form a group of variable under a latent factor heading consistent with "**Inequality**". Coinciding with the third rotated PC axis, the third latent factor account for most of the explanatory power of Rent, Death and ecks and could be dubbed "**Unrest**".

[i] *SS loading denotes the "sum of squares of factors' loadings", or the proportion of variance explained by factors in the rotated basis.*





5. Interpret scores of individuals projected on rotated components

We set out to characterize the previous rotated variable projections on the 3 significant dimentsions. Results on the right show variables correlations with rotated PC axes denoted Dim1, Dim2 and Dim3, only for p-values smaller than 5e-2. Such p-values indicate that we can reject the null hypothesis at a confidence level of 95%. In the present case H_0 posits that the correlation coefficient between the projected variable and the principal dimension is zero.

As already mentionned correlation results are only shown for variables for which we reject the null hypothesis. A special mention must be made for variable ecks, which boasts a good correlation with Dim3 and a mediocre one with Dim1 (greyed row). To note is the fact that Rent is a variable significantly correlated only with the PC3 axis and the variable Death.

$Dim.1$quanti	**correlation**	**p.value**
Laboagr	0.94	5.03E-22
demo	0.78	9.13E-11
ecks	0.41	4.40E-03
Gnpr	-0.91	4.56E-19
$Dim.2$quanti	**correlation**	**p.value**
Gini	0.92	1.65E-20
farm	0.87	2.97E-15
Instab	0.63	2.54E-06
$Dim.3$quanti	**correlation**	**p.value**
Death	0.83	4.15E-13
Rent	0.71	2.19E-08
ecks	0.71	2.31E-08

Quantitative results summarized here are in very good agreement with the conclusions we derived previosuly and independently from a purely visual examination of the plot of projected variables in a rotated basis of PC1-2 (top right plot). This opens the possibility of extracting features and reducing dimensionality of the analyzed data set, by introducing latent variables, previously identified as Wealth, Inequality and Unrest.

6-7-8. Read PCA_quetaltecaen data and symmetrize the matrix; build a dissimilarity matrix

We read the file, stored in `/Data/PCA_quetaltecaen.csv` with encoding option UTF-8. We keep the first column as row names and symmetrize the matrix by averaging over the sum of the matrix and its transpose: `s<-(s + t(s))/2`

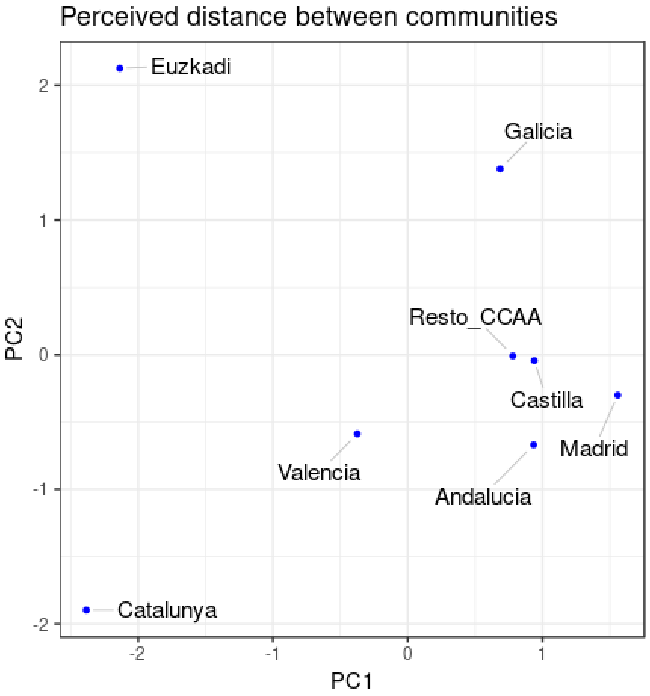
We calculate the dissimilarity matrix (`dis_qttc`) and list dissimilarity self-perceptions (diagonal elements) within Spanish regions:

Andalusia	Castilla	Catalonia	Valencia	Galicia	Madrid	Basque	Others
0.7	2.16	1.29	1.45	1.02	1.33	1.17	3.8

9-10. Conduct the PCA analysis and produce the scatter plot of Euclidian distances in PC1-2

```
    dis_data ← cmdscale(dis_qttc, k=2, eig=T, add=T)
```

Some eigenvalues were negative as a result of the PCA and so to avoid a non Euclidian distance matrix, we added the option "`add=T`" to the `cmdscale()` method above. A minimal additive constant *c* was thus computed such that the dissimilarities `d[i,j]+c` become Euclidean (within rounding errors) and can be represented in k=2 dimensions.



The most striking aspect of the above plot is how the Basques and Catalans are perceived as removed and different from the rest of Spanish regions. They are also the most distant from one another along PC2, contrasting with their proximity along PC1. To a lesser degree, Galicians are also perceived as being "apart" from the rest of spanish regions. As is the case for the Basque region and Catalonia, Galicia is a region characterized by a strong cultural identity and a specific language. By contrast the rest of regions all share castillan as their *prima lingua*.

We venture the hypothesis that the 1st PC direction (PC1) can be construed in terms of a latent factor "political ideology", whereas PC2 seems to be rooted in a more cultural measure of distance, whose best expression is perhaps the "linguistic identity". To begin to refine those preliminary conclusions about latent factors, we would need to have access to the semantics of the questionnaire submitted to the populations of those various regions.

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.