MULTIVARIATE ANALYSIS
PRINCIPAL COMPONENT ANALYSIS

Authors

Goktug Cengiz
Henry Qiu

Date
21/03/2019

This homework consist in the practice of Principal Component Analysis, there are some requirements that are splitted into different main parts in our report. These parts are: the imputation of the missing values, defining a own procedure of PCA analysis, doing the PCA analysis using the library "FactoMineR", and applying several algorithms such as NIPALS and Varimax rotation, at the end there will be a conclusion taken after doing this homework.

For each one of these parts we will provide a brief explanation of the procedure (the detailed implementation will be provided in the the submitted script), the results obtained and the answers to the corresponding questions in the statement. It is necessary to mention that plots will appear within the text to facilitate the interpretation and captures of some results that are not essentially needed to follow the explanations will appear in the annex.

Before starting, we have to threat all the missing values of the dataset so that the the PCA analysis make sense and gives us significant results, to deal with this problem we will do the imputation of missing values. We used Random Forest in this project for imputation. In Random Forest, the basic idea is to do a quick replacement of missing data and then iteratively improve the missing imputation using proximity. In order to do that, we used "missForest" package in R.

1. Defining the X matrix

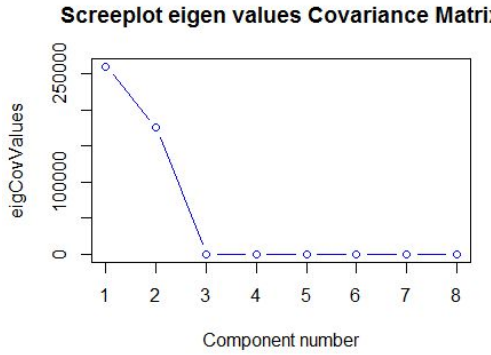
First of all, we need to create the X matrix from our dataset, the X matrix is defined only by the continuous variables, so in our case all our variables are numerical, except for the variable DEMO, that is a categorical variable, all the other variables are continuous, for this reason, we will use all the variables except DEMO to compute the X matrix. A function is provided to do this task and the returned object corresponds to the X matrix. In our case X has size 47 x 8.

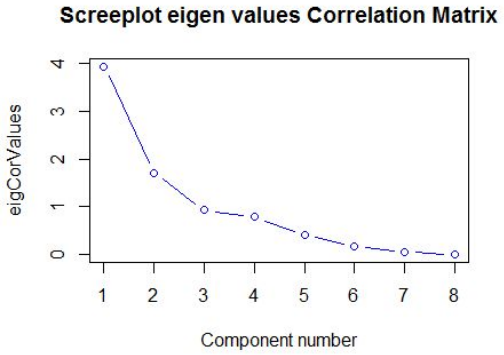
2. PCA analysis with own procedure

(a.)We will like to create weights for the individuals that can be uniformly distributed or not, to this aim, we created a function that with a boolean parameter returns a vector of weights of the size of the number of individuals.

(b,c,d,e)We computed the matrix of weights N, the centroid G of individuals, the covariance matrix, the standardized and the correlation matrix computed with the standardized matrix.

(f.)After diagonalizing the covariance and correlation matrices, we have obtained the following screeplots for their corresponding eigenvalues:



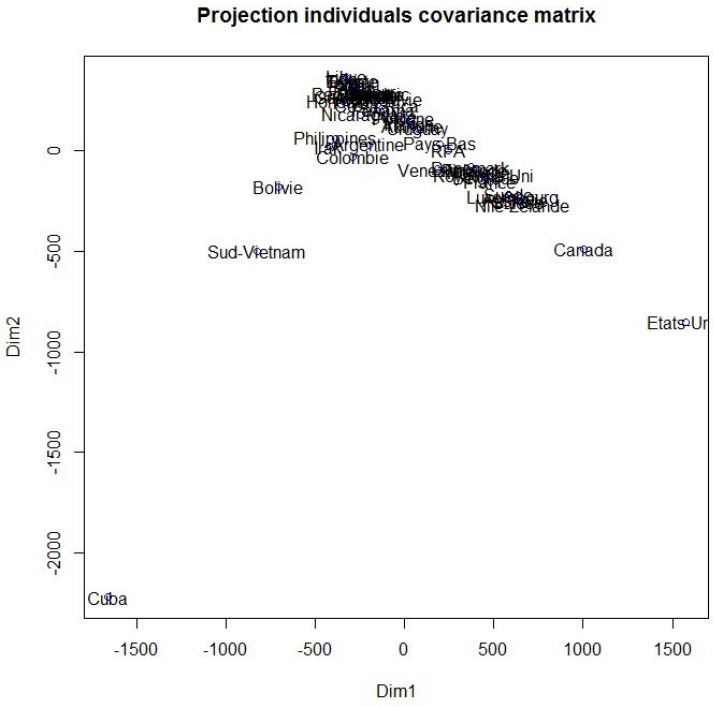


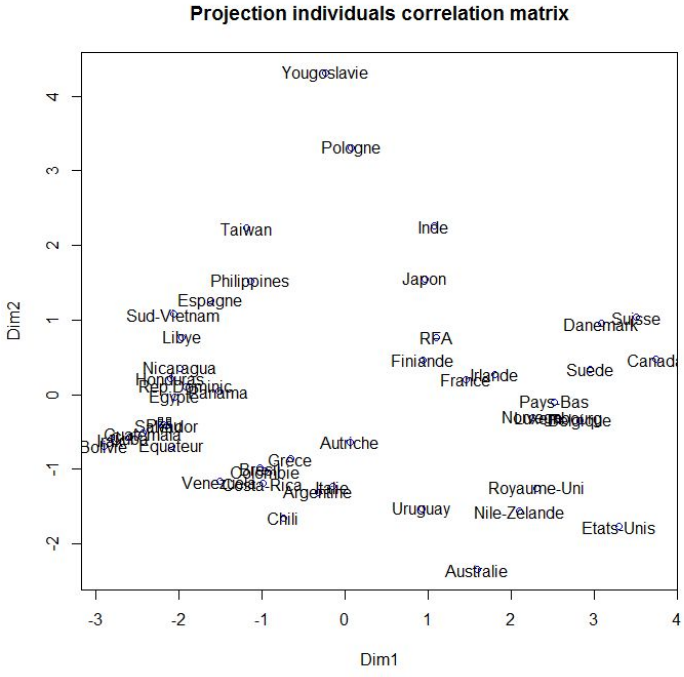
Comparing the results we can see that using the standardized matrix to compute the correlation matrix gives more compact eigenvalues, in this way the difference between the components are not huge, and that is why the plot look much more fancy than the eigenvalues of the covariance matrix.

We have to define the number of significant dimensions, so as to do that, we decide based of the information provided by the screeplots, regarding to last elbow rule, the component where the difference of pending is bigger means the components provides more information, in our case it is difficult to decide between the second or the third component. That's why we applied another method: the kaiser rule, where we are considering only the components with a variance greater than the average variance. In this case we obtained as a result 2, that means that we will take into account the components 1 and 2.

We have computed the proportion of the total variance explained by the principal component in terms of percentage, and we can see in the results that the components 1 and 2 have the highest proportion as we expected. We also calculated the cumulative percentage, we can see that the quantity of information retained in the significant dimensions we have chosen are 99.91% for the eigenvalues of the covariance matrix and 70.74% for the correlation matrix. In the first one we can see that almost all the information is concentrated in the first and the second dimension. We can see the results in the Figure 1.

(g.h.i.j)We have done the projection of the individuals with the eigenvectors of the covariance matrix and with the correlation matrix and these are the results we have obtained:





(k.)We can observe that in case of the use of non normalized matrix, the values seems to be concentrated, but that is because we have an outlier value that is Cuba which is located far away from the others, we can see that the results are much more fancy using normalized matrix, because in this dataset,without normalising some variables that are important can dominate over the others.

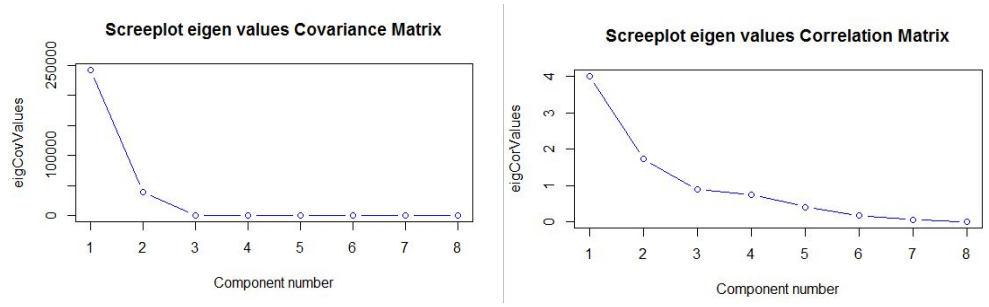
(l.) We can see the results in Figure 2. We can obtain a conclusion that the variables Rent and ecks are determine more the first dimension, in case of the second dimension, it is more determined by the variable Gini.

3. Set weight of Cuba to 0

(a.)Now we have to set the weight of Cuba to 0, so that the results of the analysis make sense, we cannot assign the weight to exactly 0, but a small number nearly 0, we will call this value Epsilon that will have value 0.001. Excepting this individual, the weight of the other individuals will be equally distributed to see the comparison of the results.

(b.c.d.e.)The alteration of the weight matrix doesn't affect the centroid, the centered matrix or the standardized matrix, using the same matrices we compute the new covariance and correlation matrices.

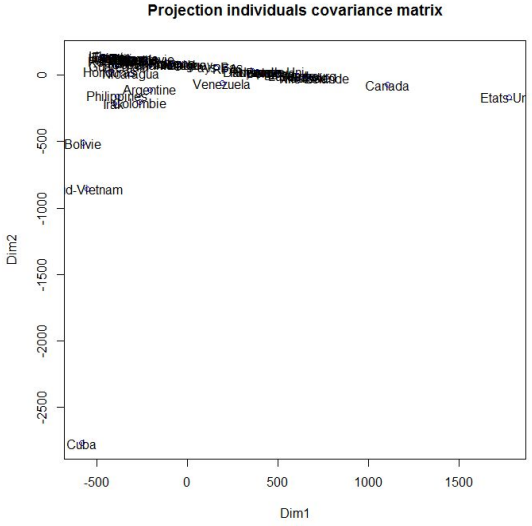
(f.)These are the screeplot of the eigenvalues with the covariance matrix and the correlation matrix.

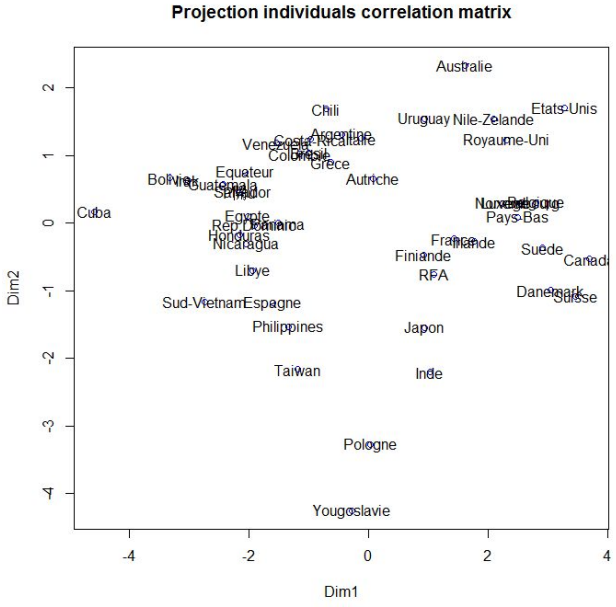


The proportion of the total variance explained by the principal component are shown in the Figure 3. The retained information of the principal components chosen are 99.86% for the eigenvalues of the covariance matrix and 71.66% for the correlation matrix.

(k.)Like before, the significant components are still the two first dimensions for the same reason mentioned before.

(g.h.i.j)After doing the projections of individuals we have obtained the following plots of individuals:



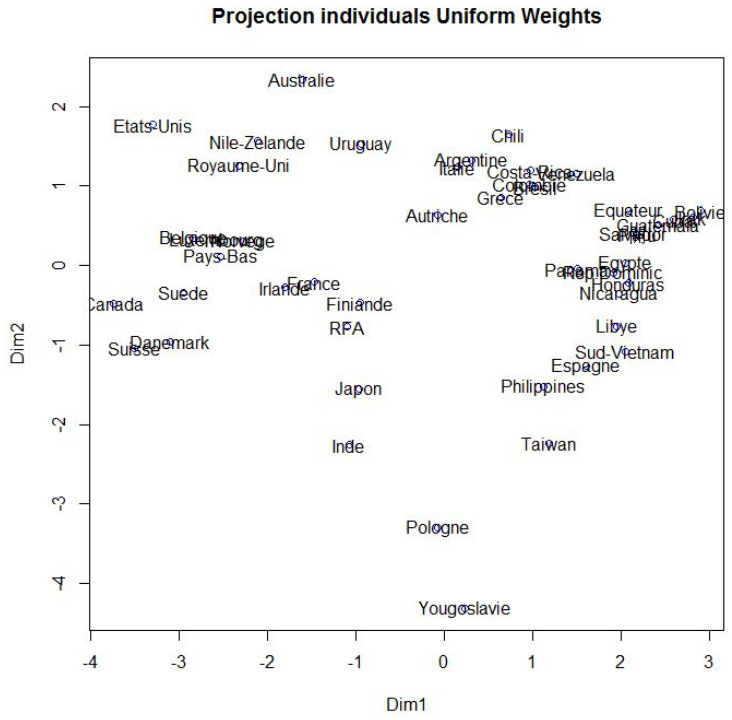


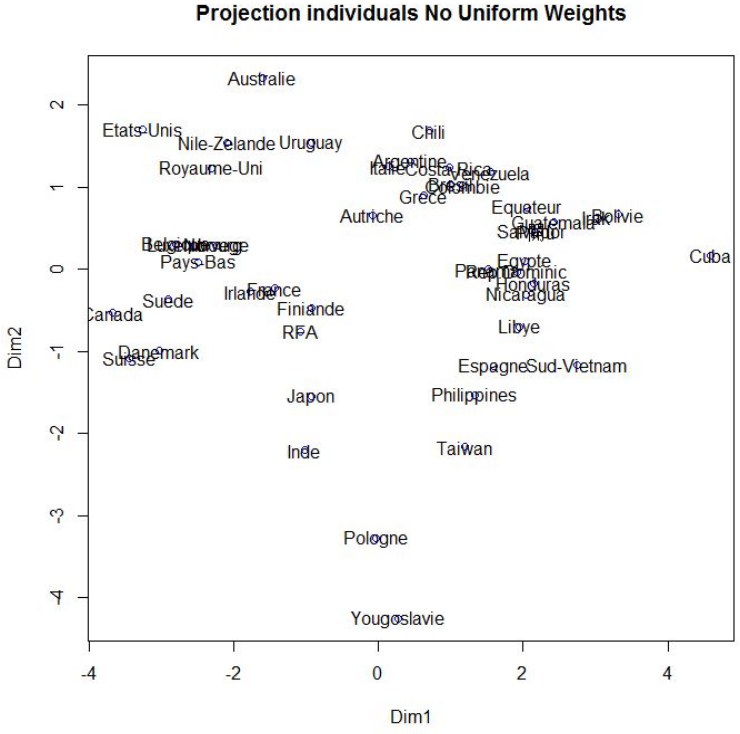
The results of analysing the correlations between the significant principal components are shown in the Figure 4. We can see that in the correlation matrix, the dimensions of the dimension 1 and dimension 2 are very high, nearly to 0.99, we can say that the two approaches are very similar. But now the outlier Cube is better represented so we can distinct it from other data.

4. Correlation between considering Cuba as outlier

We can see in the Figure5 that the correlation between both is quite high, near to 0.99, that means that they are the outlier value does not affect too much. We could say that PCA is not very sensible to outliers. However, this improvement is reflected on the quality of the plot, so that having no uniform weights seems to be a better option if we identify the outliers.

5. PCA with FactoMineR





(5.)We can see in the Figure 6 that the two plots are very related, because the value of their correlation are nearly 1. That is a good indication that our PCA method is not working well.

(6,7,8,9.)We can see in Figure 7 the answer of these questions. To compute the representation in the factorial plane we use the metric “cos2” and for the influence of formation of the components we use the metric “contrib”.

Conclusion

With these we can conclude that for the Russet dataset the normalized euclidean metric works clearly better than the non-normalized metric. First analysing the behaviour of the plots of the eigenvalues, we can see that without normalising some variables that are important dominate over the others. In case of the projections we can see that normalizing can "fix" outlier values. We can conclude that normalising the matrix will gives us better plots to facilitate the analysis specially in datasets such as Russet dataset where some variables have clearly much more importance than others.

It is also clear than using no uniform weighs can helps us to fix the outliers, so that their values are correctly represented in the way that they are different to all the other values, but at the same time it doesn't "runs out" from the space representation of the points, making the plot weird and difficult to understand.

We can also see that the PCA computed by FactoMineR is very similar to our own method, we realised that the metric used is "coord".

Annex

Figure 1

```
> proportCov
[1]  5.963714e+01  4.026955e+01  5.015200e-02  3.836634e-02  3.819723e-03  8.651984e-04  1.083126e-04 -1.908393e-15
> cumProportCov
[1]  59.63714  99.90669  99.95684  99.99521  99.99903  99.99989 100.00000 100.00000
> proportCor
[1]  4.928725e+01  2.145448e+01  1.162471e+01  9.787975e+00  4.979735e+00  2.210080e+00  6.557701e-01 -2.775558e-15
> cumProportCor
[1]  49.28725  70.74173  82.36644  92.15442  97.13415  99.34423 100.00000 100.00000
```

Figure 2

```
> corVarPr
              [,1]        [,2]
Gini     -0.5887203 -0.75054888
farm     -0.6446351 -0.66539101
Rent     -0.9126589  0.27207123
Gnpr      0.8243415 -0.24572319
Laboagr -0.8411616  0.31183225
Instab   -0.1136472 -0.63371584
ecks     -0.9126589  0.27207123
Death    -0.3390187 -0.05504057
```

Figure 3

```
> proportCov
[1]  8.615516e+01  1.370109e+01  7.858839e-02  5.761879e-02  6.017963e-03  1.355837e-03  1.707594e-04 -1.769242e-15
> proportCor
[1]  5.009250e+01  2.157027e+01  1.097997e+01  9.426475e+00  5.105731e+00  2.168643e+00  6.564097e-01 -6.938894e-16
> cumProportCov
[1]  86.15516  99.85625  99.93484  99.99246  99.99847  99.99983 100.00000 100.00000
> cumProportCor
[1]  50.09250  71.66277  82.64274  92.06922  97.17495  99.34359 100.00000 100.00000
```

Figure 4

```
> cor(PsiUniform,PsiNoUniform)
             [,1]        [,2]
[1,] 0.989920463 -0.01230972
[2,] 0.001120656 -0.99847918
```

Figure 5

```
> cor(PsiUniform,PsiNoUniform)
             [,1]        [,2]
[1,] 0.989920463 -0.01230972
[2,] 0.001120656 -0.99847918
```

Figure 6

```
> cor(PCAUniform$ind$coord[,1:dimensions],PsiUniform)
                [,1]          [,2]
Dim.1 -1.000000e+00  2.548649e-16
Dim.2  4.010457e-17 -1.000000e+00
> cor(PCANoUniform$ind$coord[,1:dimensions],PsiNoUniform)
               [,1]        [,2]
Dim.1 -1.000000000 0.005764097
Dim.2 -0.005764097 1.000000000
```

Figure 7

```
> sort(PCAUniform$ind$cos2[,1],decreasing = TRUE)[1]
Luxembourg
  0.963608
> sort(PCAUniform$ind$cos2[,2],decreasing = TRUE)[n]
      Egypte
0.0001703938
> sort(PCAUniform$ind$contrib[,1],decreasing = TRUE)[1:3]
    Canada     Suisse Etats-Unis
  7.555265   6.644661   5.860709
> sort(PCAUniform$ind$contrib[,2],decreasing = TRUE)[1:3]
Yougoslavie     Pologne   Australie
  23.113452  13.531138   6.819357
> sort(PCAUniform$var$cos2[,1],decreasing = TRUE)[1]
     ecks
0.8329463
> sort(PCAUniform$var$cos2[,2],decreasing = TRUE)[8]
      Death
0.003029464
> sort(PCAUniform$var$contrib[,1],decreasing = TRUE)[1:3]
    ecks     Rent  Laboagr
21.12479 21.12479 17.94462
> sort(PCAUniform$var$contrib[,2],decreasing = TRUE)[1:3]
    Gini     farm   Instab
32.82085 25.79561 23.39812
```