Exercise PCA

INSA Toulouse, 4 ModIA, Olivier Roustant

September 11, 2023

1 Data exploration, principal component analysis

We consider a dataset provided by the World Health Organization. For 133 countries, 21 variables are given for the year 2014, corresponding to life expectancy, vaccination, mortality, economical and social factors.

1. Figure 1 shows the boxplot of the data, without transformation ('raw' data), after scaling, and after both transformation and scaling. Comment these boxplots.
   For the raw data, we can see that one of the variable has a much larger variance than the other ones, GDP, and is very dissymetric. On the scaled raw data, we can observe that some variables are very dissymetric (GDP still, population, etc.). On the scaled transformed data, the distributions are now much more symmetric, unimodals, except some of them such as BMI.

2. In order to better investigate the interest of transforming the variables, we focus on the two variables 'Life.expectancy' and 'GDP'. Figure 2 shows the histograms of the variables (diagonal plots) and a scatterplot with or without transformation (for GDP). What is the effect of the transformation used on the distribution of GDP? What transformation could have been used? Explain why it can be useful to consider the transformed variable when the aim is to predict the life expectancy with the other variables. In particular, what can you say of the hypothesis of the linear regression of LifeEx versus the transformed variable GDPt?
   The effect of the transformation is to make the distribution more symmetric. Here a log could (and was) have been used. Interestingly here, LifeEx can be explained linearly by the transformed GDP variable, whereas the relationship between LifeEx and GDP is highly non-linear. Finally, the homoscedasticity assumption (constant variance) and seems to be rather well satisfied here. The fact that the normality assumption is satisfied by the regressor (GDPt) and the response (LifeEx) is a good thing, even if we need to check the normality of the residuals.

3. Why doing a PCA on the raw data is not a good idea here? What would be the result of such analysis?

   PCA is a variance decomposition. Here GDP has a much larger variance than the others. Thus, the first axis will explain more than 95% of the variance, and that's all that we will see. This will not bring any new information than the fact that GDP has the largest variance...

4. Figure 3 shows some results of a PCA up to dimension 3. Which variables can/cannot be used to interpret axis 1 and 2? Interpret these two axis, by writing explicitely each principal component as a linear combination of the most relevant variables (up to a multiplicative factor). Same question for axis 3. For that axis, check your answer with the graph of individuals.
   Only the variables which are closed to the border of the circle can be used, meaning that their projection is very closed to the representation space. To interpret axis 1, look among these variables those which have the strongest correlation, here Life.expectancy on the right, and mortality (Adult.Mortality, infant.deaths, etc) on the left $\rightarrow$ "life" axis. Up to a multiplicative factor, for these variables, we have

$$\begin{aligned} PC_1 \quad \propto \quad & 0.8(Life.expectancy + GDP + Schooling) \\ & -0.8(Adult.Mortality + infant.deaths + under.five.deaths) + ... \end{aligned}$$

   Second axis: mostly explained by diseases (Polio, Diphteria, Hepatite B).

$$PC_2 \quad \propto \quad 0.8(HepatisB + Diphteria + Polio) + ...$$

   Third axis: Population and Measles, mainly. For the population, confirmed by the individuals (China, India on the top, Bhoutan at the bottom).

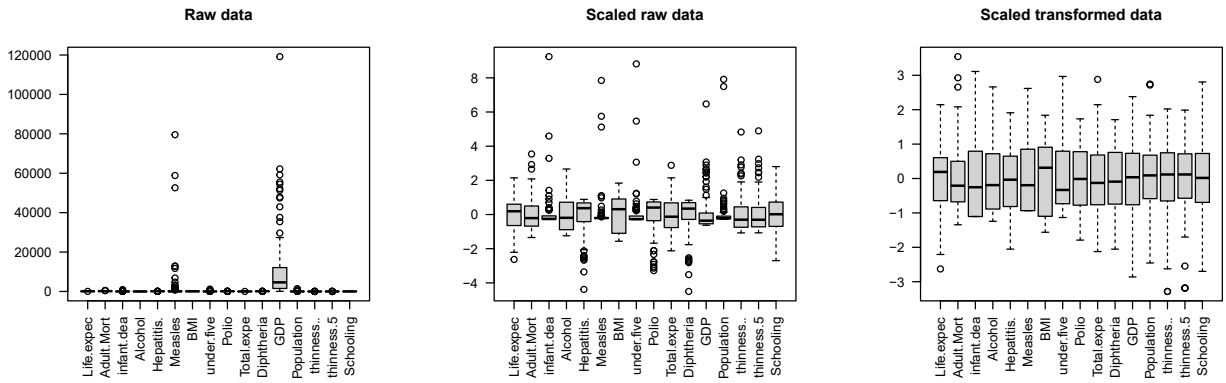

Figure 1: Boxplot of the dataset. Left: original data; Middle: scaled data; Right: transformed and scaled.

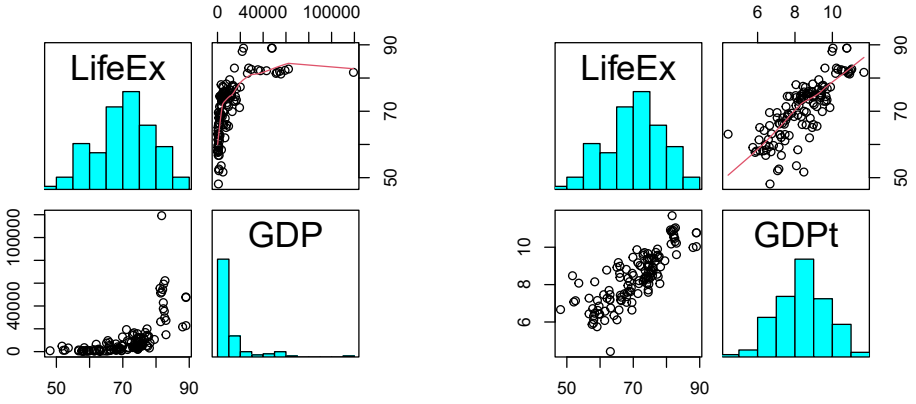

Figure 2: Scatterplots of variables 'Life expectancy' and 'GDP'. Left: raw (untransformed) data. Right: a transformation of GDP has been used.

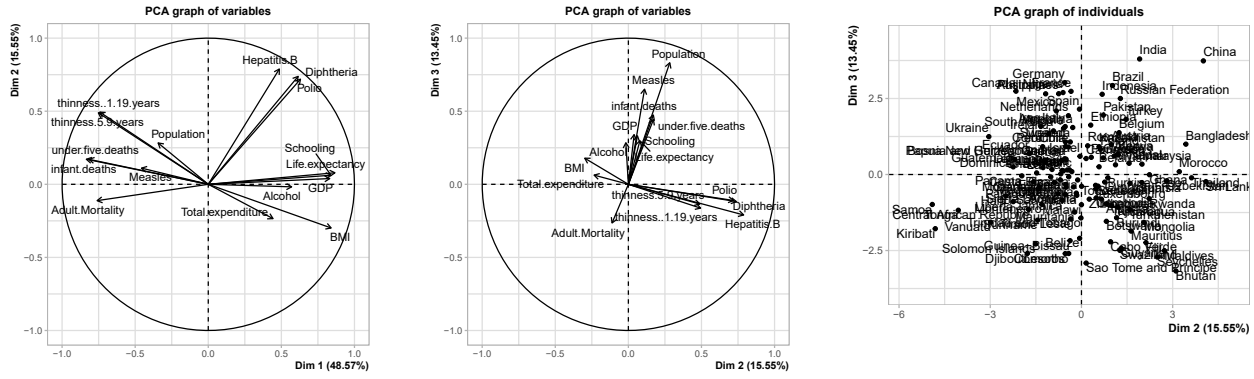

Figure 3: PCA for the Life Expectancy data set. Left: axis (1, 2); Middle and right: axis (2, 3).