Principal Component Analysis: Air pollution data

T Shapla

10/25/2023

#Principal Component Analysis (PCA)

## Example Air Pollution data

Below we consider ‘Air Pollution’ example. The accompanying data is provided in Table 1.5, page 39 of the textbook (Johnson), and in AirPollution data in Canvas Data module.

In the Air Pollution data, there are 42 measurements on 7 air-pollution variables recorded at 12:00 pm in the Los Angeles area on different days. The variables are given below. For the sake of simplicity, we assume that all the variables have same unit of measurement.

1. Wind
2. Solar radiation
3. CO (Carbon monoxide)
4. NO (Nitrogen oxide)
5. NO2 (Nitrogen dioxide)
6. O3 (Ozone)
7. HC (Hydrogen carbide)

Read the data in R, assign variable names, and answer the following questions:

1. To perform a PCA for the above data, should you use a sample covariance matrix or correlation matrix and Why? Using the appropriate matrix, compute and write down the first and second principal components, and .
2. Draw a screeplot and provide your graph below. Based on the graph, decide how many principal components you will consider in your study.
3. How many components should you consider to account for at least 85% of the total variation in all the variables?
4. Based on using the average of eigenvalues as a cutoff, how many principal components you will consider in your study?

e.. Find the sample variance of each variable and explain how the relatively larger variances affect the principal component loadings.