Dimension Reduction

Tutorial 1 and 2

Anastasios Panagiotelis

November 16-23, 2022

R Markdown

World Bank Data

Carry out principal components analysis on the World Bank data (`WorldBankClean.csv`) and answer the following questions. To answer the questions run the command `?prcomp` to see the help documentation of the function used to create principal components.

1. Do you scale the data before applying principal components? Why or why not?
2. What is the standard deviation of the first principal component?
3. How many principal components are needed to explain at least 90% of the total variance?
4. Plot the first two principal components of the data as a scatterplot. Use the country abbreviation rather than points.
5. What are the loadings (weights) on the first principal component of infant mortality rate (SP.DYN.IMRT.IN- variable 78) and number of pupils in primary education (SE.PRM.ENRL - variable 123).
6. What are the loadings (weights) on the second principal component of infant mortality rate (SP.DYN.IMRT.IN- variable 78) and number of pupils in primary education (SE.PRM.ENRL - variable 123).
7. Do these results make sense in light of our interpretation of the first PC measuring level of development and the second PC measuring the size of a country?
8. Run kernel PCA using the `dimRed` package. Use a polynomial kernel with a degree of 3, a scale of 0.001 and an offset of 1. Use `?kPCA-class` and `?kpca` to consult the help documentation of the function in the `dimRed` package and the original function from the `kernlab` package that it wraps around.

Irish Smart Meter data

1. Using the household with ID=4669 for the Irish smart meter data, find the location and scale parameter from fitting electricity demand for each time of week with a lognormal distribution

HINT 1: You only need to take the log of the data then find the mean and standard deviation, but you will need to remove a small number of observations for which demand is 0.

HINT 2: The `filter` function can be used to remove zeros, the `mutate` to take the log and the `group_by` and `summarise` functions can be used to find means and standard deviations by the time of week (`tow`) variable. These are all functions from the `tidyverse`.

2. Compute the Jensen Shannon distance (JSD) between all pairs. Remember that the JSD is the square root of the average of the Kullback Leibler divergence from P to Q and the Kullback Leibler divergence from Q to P. The Kullback Leibler divergence from P to Q for a lognormal distribution is given by.

$$KL(P||Q) = \log \sigma_Q^2 - \log(\sigma_P^2) + \frac{\sigma_P^2 + (\mu_P - \mu_Q)^2}{2\sigma_Q^2} - 0.5$$

3. Construct a plot similar to the one on slide 20 of Lecture 3. A simple scatterplot is sufficient, you do not need to replicate the coloring of the plot in the slides.

4. There are six points with a very high value on the first coordinate and a low value on the second coordinate. What do the points in this underlying group have in common. HINT: To investigate use the `arrange` function

Bonus Questions

1. The second principal component is the linear combination of weights with maximal variance while also being uncorrelated with the first principal component. Prove that the weights of this combination are given by the eigenvector corresponding to the second largest eigenvalue.

2. For the polynomial kernel

$$K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + 1)^2$$

in the case where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, show that a feature map with this kernel is given by

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \\ 1 \end{pmatrix}$$