May 2024

Assignment 5: Principal Component Analysis

In both exercises below, use linear algebra built-in functions in Python in your code, such as built-in functions for SVD. Do not use higher-level libraries for PCA (the focus in this course is on the linear algebra behind PCA).

1. Use the same tiny data matrix as in previous assignment (the movie matrix):

$$A = \begin{array}{ccccc} \text{Drama 1} & \text{Drama 2} & \text{SciFy} & \text{Documentary} & \\ 5 & 5 & 0 & 4 & \text{Ali} \\ 1 & 1 & 5 & 0 & \text{Beatrix} \\ 3 & 2 & 0 & 4 & \text{Elsa} \\ 5 & 3 & 0 & 5 & \text{Johan} \\ 0 & 0 & 4 & 0 & \text{Chandra} \end{array}$$

   Investigate the correlation between the variables (the movies, the columns) and find the direction that covers most of the covariance in the data. Do this through principal component analysis. Write a python code, and try to answer the below questions. Include a plot of the singular values in your code.

   a. Calculate the eigenvalues without explicitly forming the covariance matrix, and also each eigenvalues fraction of the total variance. How many percent of the total variance is covered by the first eigenvalue?

   b. Calculate the principal components. Try to interpret the principal component related to the largest eigenvalue. What do the vector mean in in terms of groups in the data that are highly related, and groups that are not so much related?

2. The Framingham Heart study is large health study, started in 1948 and is still ongoing. The study has been working with different generations of cohorts over time, and they have more recently also looked at different food regimes and how they affect the cardiovascular health. Early on, they identified major risk factor for cardiovascular disease (CVD), such as blood pressure, triglyceride and cholesterol level, age etc. Many of the diet recommendation that are believed to reduce risk for heart disease are based the Framingham study. You can read more about the study here: https://www.framinghamheartstudy.org/fhs-about/.

   In the file **`framingham_heart_disease.csv`** you'll find data from one of the early cohorts. The data contains 4238 samples and 16 variables (stored in the columns). Some of the columns are just groups (male/female, current smoker etc) that are either true/false (or 0/1). The columns labelled **age**, **cigsPerDay**, **totChol**, **sysBP**, **diaBP**, **BMI**, **heartrate** contain the variables we will look at here, which makes 7 variables (you can remove the **education** variable here).

   Find the three dominating principal components in the data, using the 7 variables

above, and calculate their fraction of the total variance. Plot the principal components in 3D, grouped after **prevalentHyp** (meaning prevalent hypertension). Also, plot the principal components in 2D (principal component 1 versus 2, principal component 1 versus 3), also grouped after prevalent hypertension.
Try to interpret the results above. What does it mean?
Also, look at the first eigenvector (it's just 7 variables, so you can actually look at it) and try to interpret what it tells us.