ST437/537 – HW #03

Arnab Maity

Due date: January 31, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework before coming to class. Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- **DO NOT** give printouts of whole dataset or matrices. Present only the relevant output when answering a question.

Problems

Solve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

DO NOT give printouts of whole dataset or matrices. Present only the relevant output/graphs when answering a question.

1. (15 points) Examine the `USJudgeRatings` data in the datasets library. This dataset contains the ratings of 43 US Superior Court judges by attorneys. Each of the judges is evaluated on each of 12 attributes such as demeanor, preparation for trial, sound rulings, and the number of contacts each attorney had with that judge. See the R help file for more information on this dataset.

a. Examine the dataset (especially the relationship among variables) using 1 or 2 plots we discussed in class.
b. Perform PCA on the dataset, and retain the first few PCs that capture *at least* 90% of total variability. Report the results.
c. Interpret the loading of the PCs you retained. Does your interpretation match with your findings in part (a)?

2. (20 points) The dataset `Harmon23.cor` in the datasets package is a correlation matrix of eight physical measurements made on 305 girls between the ages of 7 and 17.

a. Provide a plot to visualize this matrix, and comment on any patterns you see.

b. We want to perform PCA on this matrix (`Harman23.cor$cov`), and retain the first two PCs. How can we do this? [Note: this is not the actual dataset, you only have the correlation matrix of the variables.]

c. How much of the total variation is captured by each of the first two PCs? How much of the total variation is captured by the two PCs together?

d. Interpret the first two PCs.

3. (15 points) Consider a principal components analysis on the [cancer dataset] (cancer.txt) appearing in Table 1.4 of Zeltermann: Applied Multivariate Statistics with R. This table lists rates of individual cancer types (and overall rates) for each of the 50 US states plus DC.

a. Examine the standard deviations of each cancer type, including the overall rates.

b. Do you think it is more appropriate to examine the covariance (i.e., use the data matrix as it is) or the correlation (i.e., first standardize the data and then compute covariance) in a principal components analysis of this data? Explain.

c. Perform a principal components analysis on this dataset and report the results.

4. (20 points) Consider the `heptathlon` data in the `HSAUR3` package. See `?heptathlon` for details about the dataset.

a. Take only the first seven columns representing the seven events. Notice that in the events high jump, long jump, shot and javelin, larger values indicate better performance. But for the other three events (200m, 800m, and hurdles) smaller values indicate better performance. To help with interpretation, transform the data of the latter three events as “newx <- max(x) - x” so that for all the variables larger values indicate better performance. Visualize the correlation matrix and comment on any pattern you see.

b. Perform PCA on the new dataset. Summarize and interpret the results, especially the first two PCs. Note that you might need to standardize the data.

c. Compute the PC scores for the first two PCs and create a scatterplot. Do you see any pattern? If yes, investigate further and comment on your findings.

d. The last column of the `heptathlon` dataset provides the official scores given to the athletes for the event. Note that your PCA did not involve the score information at all. Consider PC1 (a summary of the performance). Plot the official scores versus your PC1 scores in a scatterplot. Do you think your summary of performance, PC1, aligns with the official scores? Comment on your findings.