STAT 5353: R Assignments
Due by Wednesday, Dec 13, 2017 10:00 AM

Read the following rules carefully:

A. There are 3 questions in this assignment. You may answer all of them. **I shall exclude the one with lowest score and consider the remaining two.**

B. For each question, separate your answer in two parts as follows:

1. Part 1 contains the answers (more on that below) without any R code. It must have all numbers/values and appropriate explanations in it. You can handwrite it or type it. You can submit in hard copy or email – whichever you prefer.

2. Part 2 contains only the R code and nothing else. **Any output/numbers/explanations written inside the R Code will NOT be graded**. **This part must be emailed.**

C. Your submission of both Part 1 and Part 2 must be complete by 10 AM on 12/13/2017 - the completion of scheduled in-class final exam.

D. In Part 1, do not **only** write the numbers that you get from R. You need to explain the necessary details from the notes. Whenever you are using any particular formula or technique you have to first mention that and then give the output corresponding to that.

E. **Any correct answer, without appropriate description of method, is not going to get any point.**

R Assignment 01

Q1. Consider the Pollution dataset. The dataset consists of measurement of several pollutants in different factories. Each row represents data from one factory and each column represents data from one pollutant. Now, answer the following questions

[1+2+4+2+2+1+3=15 points]

(a) How many factories are there in the dataset? How many pollutants are there?

(b) Use Parallel Analysis to choose the number of factors k using mean as summary criterion. Use 250 chains. Report the comparison table and your selected number of factors.

(c) Construct the factor loading matrix using PCA method and then do a varimax rotation. Report the improvement for each factor by computing the **ratio** between initial and final variances. Which factor showed maximum improvement due to rotation? Which factor showed minimum improvement due to rotation? Also, report the overall improvement due to varimax rotation by taking the **ratio** before and after rotation.

(d) Now, find the number of pollutants for which at least 90% variability was explained by factors. What percentage of variability of all pollutants were explained by 2nd factor?

(e) Find the factor score for 3rd Factory.

(f) Identify the pollutant whose variability was best explained by the factors.

(g) Suppose, I decide to use 10 factors. Then compute L as in (c) (This time, you do NOT need to discuss the improvements for varimax). Now, recompute your answers for (d) and (f).

R Assignment 02

Q2. Consider the genetic dataset. The dataset consists of gene expression measurements collected from different tissues. Each row represents data from one tissue. Each column represents data from one specific gene. Now, answer the following questions

[1+2+4+2+2+1+3=15 points]

(a) For this problem, find if $n < p$ or $n > p$?

(b) Use Parallel Analysis to choose the number of factors k using 95% quantile as summary criterion. Use 150 chains. Report the comparison table and your selected number of factors.

(c) Construct the factor loading matrix using PCA method and then do a varimax rotation. Report the improvement for each factor by computing the **difference** between initial and final variances. Which factor showed maximum improvement due to rotation? Which factor showed minimum improvement due to rotation? Also, report the overall improvement due to varimax rotation by taking the **difference** before and after rotation.

(d) Now, find the number of genes for which at most 25% variability was explained by factors. Identify the gene whose variability was worst explained by the factors.

(e) Find the correlation between 3^{rd} and 17^{th} genes. Find the correlation between 5^{th} gene and 3^{rd} factor.

(f) Among gene 14 and gene 40, which is better explained by factors? Justify.

(g) Suppose, I decide to use 8 factors. Then compute L as in (c) (This time, you do NOT need to discuss the improvements for varimax). Now, recompute your answers for (d) and (f).

R Assignment 03

Q3. For this question, refer to the compressed file Face Image Data. It contains a folder named "yalefaces_train" and another folder "yalefaces_test". Use the yalefaces_train folder to do a PCA as follows: [2+1+4+4+4=15 points]

(a) What are the values of n and p?

(b) Plot the "average face".

(c) Do a PCA allowing 20% loss of information. What is the value of k? Plot the third eigenface.

(d) Now, choose any one image from yalefaces_test folder (**Every student's choice should be independent of any other student's choice**) Reconstruct your chosen image using PCA from part (c). Plot the original image and reconstructed image.

(e) Repeat (d) using the same image that you have chosen before but allowing 6% loss of information. What is the value of k now? Plot the reconstructed image. Compare between the qualities of two reconstructions that you get now with the one you got in (d).