University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

STAT40150 Multivariate Analysis

Trimester 2 2019/2020

Assignment 3

Associate Professor R. Killick
Professor T. B. Murphy
Associate Professor I.C. Gormley*

Instructions for Candidates

Full instructions are detailed on page 2.

Instructions

- There is a total of 200 marks for this assignment and it is worth 45% of your final module mark.
- Answer all questions. Carry out all analyses in R.
- Due date: **11pm Saturday May 23rd 2020**. Please submit your assignment by uploading **ONE PDF FILE** to BrightSpace containing your answers to the questions, with the (fully commented) R code (if required) embedded within each question's solution. It should be possible for the grader of your assignment to copy and paste this code into R, thereby reproducing your work. Students may submit using R Markdown if they wish but this will gain no additional marks. **ASSIGNMENTS SUBMITTED BY EMAIL WILL NOT BE ACCEPTED AND WILL NOT BE GRADED.**
- Your assignment should include solutions to the questions posed below. These solutions may include scanned copies of written responses, typed responses, output and plots from R. All code should be fully commented within your answers to clearly illustrate how you arrived at your solution for each question. NB: where relevant if R code is not provided, marks will not be awarded.
- Any plots included should be clearly labelled.
- While discussion of the problems is encouraged plagiarism is not permitted. Anyone found to have been involved in plagiarism will score 0.
- Late assignments will be graded according to UCD's Late Submission of Coursework Policy.
- Please include as your first page in your submitted pdf a copy of the following text with your details inserted, indicating your adherence to the UCD School of Mathematics and Statistics Exam Honour Code.

  UCD School of Mathematics and Statistics Exam Honour Code.

  I confirm that I have not given aid, or sought and/or received aid for this assignment.

  Name:

  Student Number:

1. In order to derive required drug dosages Milner and Rougier (2014) recorded a number of variables on a cohort of 544 Kenyan donkeys. The data are available on Brightspace as `Donkeys.csv`.

Variable	Measurement scale
Girth	cm
Height	cm
Length	cm
Weight	kg
Age	<2, 2-5, 5-10, 10-15, 15-20, >20 (years)
BCS	Body condition score: 1 (emaciated) to 3 (healthy) to 5 (obese) in steps of 0.5.
Sex	Female, gelding, stallion.

(a) Use suitable plots to illustrate each of the variables in the donkey data set. Comment on the distributions of the variables, and on the outlying donkey. Remove the outlying donkey from the dataset.

[5 marks]

(b) A principal components analysis is employed to reduce the dimension of the donkey data. Which variables should be used in such an analysis?

[5 marks]

(c) Would you advise performing principal components analysis on the correlation matrix or covariance matrix of the appropriate donkey variables? Explain you reasoning.

[5 marks]

(d) Write **your own function** that could be used to apply principal components analysis to a multivariate data set. You should not use any inbuilt PCA functions that are available in R, but should derive the method from first principles and write **your own code** to implement the method accordingly. Your function should output objects that would be of interest to someone using your function.

[25 marks]

(e) Set the seed in R to your student number. Randomly sample 5 values between 1 and 500, and remove the corresponding donkeys from the data set. All subsequent analyses in question 1 should be conducted on this version of the dataset. From the output of the application of your own PCA function to the appropriate variables, how many principal components are required to summarise the donkey data? Use suitable plot(s) to motivate your decision.

[15 marks]

(f) Interpret the first column of the loadings matrix resulting from the application of your PCA function to the appropriate variables from the modified donkeys data (from 1(e)).

[5 marks]

(g) Plot the first principal component scores of the donkeys resulting from the application of your PCA function to the appropriate variables from the modified donkeys data (from 1(e)). Why is such a plot useful in PCA? Comment on the principal component scores in the context of the available data.

[20 marks]

(h) The jackknife is one method that could be used to validate the principal components solution. Detail in your own words how the method works. **Write your own code** to implement the method. Use your code to validate the results obtained from applying your PCA function to the appropriate variables from the modified donkeys data (from 1(e)).

[20 marks]

[Total: 100 marks]

2. Under the factor analysis model a p-dimensional observation $\mathbf{x}_i$ $(i = 1, \ldots, N)$ is modeled as

$$\mathbf{x}_i = \mu + \Lambda \mathbf{f}_i + \epsilon_i$$

where $\mathbf{f}_i \sim MVN_q(\mathbf{0}, \mathbf{I})$ and $\epsilon_i \sim MVN_p(\mathbf{0}, \Psi)$, where $\mathbf{f}_i$ and ϵ_i are assumed independent and $q \ll p$.

(a) Show that $\mathbf{x}_i \sim MVN_p(\mu, \Lambda\Lambda^T + \Psi)$ detailing any assumptions required.

[30 marks]

(b) Describe why a factor rotation is often employed when fitting a factor analysis model.

[10 marks]

(c) Data were collected from 99 ash samples originating from different biomasses. For each ash sample the Softening Temperature (SOT) in degrees centigrade was recorded. The mass concentrations of each of eight elements (P2O5, SiO2, Fe2O3, Al2O3, CaO, MgO, Na2O, K2O) was also experimentally determined for each ash sample. The data are available on Brightspace in the file `AshData.csv`.
Set the seed in R to your student number. Randomly sample 5 values between 1 and 99, and remove the corresponding ash samples from the data set. All subsequent analyses in question 2 should be conducted on this version of the dataset.

i. Plot the mass concentration data for the ash samples. Would you advise transforming the data prior to the application of factor analysis? Explain your reasoning.

[10 marks]

ii. Apply factor analysis, employing a varimax rotation, to the (possibly transformed based on your answer to 2(c)(i)) mass concentrations of the eight elements. How many factors do you think are required to capture the correlation structure in the variables? Explain your reasoning. Interpret the first two columns of the loadings matrix.

[30 marks]

iii. Plot the ash samples' factor scores on the first latent factor, illustrating the SOT variable for each sample. Comment on any observed patterns.
[20 marks]

[Total: 100 marks]

—o0o—