STAT40150 Multivariate Analysis Assignment

Professor Claire Gormley

Trimester 2 2022/2023

• There is a total of 175 marks for this assignment and it is worth 40% of your final module mark.

• Answer all questions and carry out all analyses in R.

• Due date: 4pm Thursday March 30th 2023. Please submit your assignment by uploading (i) a pdf file to BrightSpace containing your answers to the questions, and (ii) a *fully commented* R script which clearly indicates how you obtained your answers. 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 a single pdf containing answers and code generated using R Markdown if they wish but this will gain no additional marks. Assignments submitted by email will not be graded.

• If submitting a single pdf generated using Rmarkdown, the pdf should be no longer than 20 pages. If submitting a pdf containing answers to the questions only, it should be no longer than 10 pages. Your submission should be as concise as possible, both in terms of commentary, output included and code.

• Full reasoning should be provided for any decisions made throughout the analyses.

• Late assignments will be graded according to UCD's Late Submission of Coursework Policy, as detailed on the module's Brightspace page.

• Your pdf should include solutions to the questions posed below; these solutions may include text, necessary output and plots from R. In your R script, all code should be fully commented 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, in any form, is not permitted.* Students should familiarise themselves with the plagiarism policy detailed on the module's Brightspace page.

- This assignment is based on the milk spectra data set (`Milk_MIR_Traits_data_2023.csv`) available on Brightspace. Background to these mid-infrared (MIR) spectral data extracted from milk samples, and their associated protein and technological traits, is provided in the recorded invited lecture from adjunct Assoc. Prof. Sinead McParland (see Brightspace), and in the pdf slidedeck SMcParland_LectureSlides_2023.pdf on Brightspace. Download the dataset Milk_MIR_Traits_data_2023.csv from Brightspace.

- The initial columns in the dataset contain details (i.e. covariates) of the cows which produced the milk samples and the protein and technological traits of the milk samples measured in the laboratory. The data in the final 531 columns are the MIR spectra, with the first row of these columns detailing the wavenumber (measured in cm^{-1}). The spectral values in the dataset are the absorbance values (the log10 of the reciprocal of the transmittance value). The water region has been removed.

1. Load the data set into R. Use the `set.seed` function in R to set the seed to your student number. Randomly generate a number between 1 and n (where n is the number of rows in the dataset), and delete that observation/row from the dataset. Ensure that you include the code used in this step in the R code you submit with your assignment so that your work can be reproduced. [0 marks]

2. The milk protein β Lactoglobulin B is used in the production of protein drinks. Remove from the dataset any record/observation which has a missing/NA value for β Lactoglobulin B. Then, visualise the spectra and the protein trait β Lactoglobulin B using (separate) suitable plots. Comment on the plots. Remove any observations with β Lactoglobulin B outside of 3 standard deviations from the mean of the trait. [10 marks]

3. Use hierarchical clustering and k-means clustering to determine if there are clusters of similar MIR spectra in the data. Motivate any decisions you make. Compare the hierarchical clustering and k-means clustering solutions. Comment on/explore any clustering structure you uncover, considering the data generating context. [25 marks]

4. Apply principal components analysis to the spectral data, motivating any decisions you make in the process. Plot the cumulative proportion of the variance explained by the first 10 principal components. How many principal components do you think are required to represent the spectral data? Explain your answer. [10 marks]

5. Derive the principal component scores for the milk samples from first principles (i.e., you should not use an inbuilt function such as predict(...)). Plot the principal component scores for the milk samples. Comment on any structure you observe. [15 marks]

6. Interest lies in predicting the β Lactoglobulin B trait based on the MIR spectra. Principal components regression (PCR) is one approach to doing so for such $n < p$ data. Research the principal components regression method and how it works e.g., see An Introduction to Statistical Learning with Applications in R by James et al. (2021), The Elements of Statistical Learning by Hastie et al. (2017), and/or the peer-reviewed journal article The pls Package: Principal Component and Partial Least Squares Regression in R by Mevik and Wehrens (2007).
   In your own words, write a maximum 1 page synopsis of the PCR method. Your synopsis should (i) explain the method's purpose, (ii) provide a general description of how the method works, (iii) detail any choices that need to be made when using the method and (iv) outline the advantages and disadvantages of the method. [30 marks]

7. Use the function `pcr` in the `pls` R package to use PCR to predict the β Lactoglobulin B levels from the spectra for a test set, where the test set is one third of the data. Motivate any decisions you make. [25 marks]

8. Seven milk proteins, one of which is β Lactoglobulin B, are important for the production of cheese and whey (see invited lecture slides). Here, for some records/observations the β Lactoglobulin B values are exactly 0, while there are non-zero values for the other milk proteins for the same records. Often records with such strange measurements are deleted, arguably losing information.
   Here, rather than delete these observations, the β Lactoglobulin B values of 0 could be treated as 'missing at random'. Often such missing values are imputed using e.g., the mean of the observed β Lactoglobulin B values. In the multivariate setting, matrix completion methods can be used to impute such missing at random values. (Note that matrix completion approaches are often used to power recommender systems such as Netflix.)
   One matrix completion method uses principal components analysis as detailed in section 12.3 in An Introduction to Statistical Learning with Applications in R by James et al. (2021). Read this section to understand how the method works. Write your **own code** to impute the β Lactoglobulin B values that are 0 using principal components analysis on the seven milk proteins data. You must use the function `prcomp` or `eigen` in your solution. Comment on the results you obtain. [30 marks]

9. Using PCR, predict the β Lactoglobulin B values from the MIR spectra for a test set where the training set contains:

(a) all records with an observed, non-zero value of β Lactoglobulin B.
(b) all records but where 0 values of β Lactoglobulin B are imputed using the observed mean.
(c) all records but where 0 values of β Lactoglobulin B values are imputed using principal components analysis.
   Comment on what you observe. [30 marks]