STAT40150 Multivariate Analysis Assignment

Dr. Garrett Greene

Trimester 2 2024/2025

Assignment Instructions

• **There is a total of 160 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: 5PM Friday March 28th 2024. 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.**

Dataset Details

This assignment is based on a dataset generated as part of a study to map the pressure distribution created by people lying in different postures. See for reference https://ieeexplore.ieee.org/document/7897206 .

This study employed mattresses equipped with pressure sensors in a 64x27 grid. Eight different subjects were recorded lying in a range of different positions, and a pressure map in the form of a 64x27 grid of pressure values produced for each. In total 462 such pressure maps were produced.

These pressure maps are of interest in a number of medical applications, as they can be used to assess and classify sleep patterns, aid in the diagnosis of sleep disorders and conditions such as sleep apnoea, or monitor activity of unconscious and comatose patients.

For the purposes of this project, these pressure maps were downsampled to a 16x9 grid and reshaped to give a dataset containing 144 variables, labelled V1 up to V144. Thus V1 corresponds to the pressure in the top left corner of the mattress, V16 the pressure in the bottom left, and so on, with V144 corresponding the the bottom right corner. Each of the 462 observations in our dataset contains one such pressure map.

In addition to the 144 pressure values in each observation, factor variables representing the subject ID, the Posture of the subject and the type of mattress used are also included. Furthermore, the age, height and weight of each subject was recorded in a separate file. The original dataset from which these data were derived, as well as some background information can be found here. However only the subset of data included in the *Pressure_Data.csv* file on Brightspace should be analysed.

Questions

1. Load the data set *Pressure_Data.csv* from Brightpsace into R. Use the `set.seed` function in R to set the seed to your student number. Using `R` random number functions, select a random subset of 400 observations from the dataset. In this way you will achieve an (almost certainly) unique dataset for your own analysis. 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. [5 marks]

2. The factor variables *Mattress_type*, *Position*, *Posture*, and *Subject* contain information on the type of mattress used, the Position the subject lay in, the Posture (left-side, right-side or supine) and the identity of the subject. Remove from the dataset any record/observation which has a missing/NA value for any of these variables. Then, visualise the the pressure measurement variables (V1, ... V144) using suitable plots. You may need to select subsets of the variables to allow plotting. Comment on the distribution of the pressure variables. Identify any potential problems, and deal with them using appropriate methods. Justify all choices made. [20 marks]

3. Use hierarchical clustering and k-means clustering on the pressure meaurements to determine if there are clusters of similar profiles 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. The pressure map data may be used to identify sleeping patterns, including classifying the sleeping posture of the subject. Perform a linear discriminant analysis to classify subjects by posture, using the pressure data (V1-V144). Assess how well your classifier performs using an appropriate method. Can you fit a Quadratic Disciminant Analysis to these data? Explain your answer with reference to the definition of the LDA and QDA models and the corresponding discriminant functions. [25 marks]

5. Apply principal components analysis to the pressure data, motivating any decisions you make in the process. Plot the cumulative proportion of the variance explained by the principal components. How many principal components do you think are required to represent the data? Explain your answer. Using e.g. colour plots or other visualisation method, provide some interpretation of the first two principal components. [20 marks]

6. Using a two-dimensional representation based on the first two principal components found in the previous step, produce a plot showing the linear bayes decision boundaries arising from your LDA model. (see https://rpubs.com/ZheWangDataAnalytics/DecisionBoundary for an example of how to plot a linear decision boundary). [15 marks]

7. Use PCA to reduce the dimensionality of the pressure data, and fit an LDA and QDA to classify the observations by *Subject*. Compare and comment on the performance of the models. Did you perform PCA using the covariance matrix, or the correlation matrix? Explain the significance of this choice, and comment on which choice is most appropriate in this case. Verify this with reference to the data and the performance of your classification models. [20 marks]

8. We would like to be able to estimate the Body Mass Index (BMI) of a patient based on the pressure measurements. Since the 144 pressure variables are highly correlated, regression models applied to them may be extremely overfitted. Principal components regression (PCR) is one approach to "regularising" regression for such highly correlated 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. [15 marks]

9. Use the information contained in the file *Subject_Info_Data.csv* to compute the BMI for each subject, and merge this information into the pressure dataset. Divide your data into training and test sets, and use the function `pcr` in the `pls` R package to perform PCR on the training data. Use your fitted model to predict the BMI for observations in the test set. Motivate any decisions you make and evaluate the performance of your model. [15 marks]