STAT40150 Multivariate Analysis: Assignment 2

Associate Professor Claire Gormley

Trimester 2 2019/2020

• **There is a total of 100 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: 11pm Monday, April 13th 2020 (week 11). 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 graded.**

• **Late assignments will be graded according to UCD's Late Submission of Coursework Policy.**

• **Please include your student name, student number and module code at the top of your submission.**

• **Your assignment should include solutions to the questions posed below; these solutions may include written/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.**

Question 1

Ireland is a parliamentary democracy. The National Parliament (the Oireachtas) consists of the President and two Houses: Dáil Éireann (House of Representatives) and Seanad Éireann (the Senate). The members of Dail Éireann, called Teachtaí Dála (TDs), are directly elected by the people. More information may be found at www.oireachtas.ie.

The votes of TDs are recorded and posted on www.oireachtas.ie. During the 5th session of the 32nd Dáil Éireann (7 September 2019 - 14 January 2020 inclusive) there were a number of votes in Dáil Éireann by the set of elected TDs. Data on whether each TD voted yes or not for six of these votes have been downloaded from the Oireachtas website and are posted on Brightspace in the file `32ndDail_FifthSession_BinaryVotes.Rdata`. The data record if a TD was voted yes (coded 2) or not (coded 1). The main topic of each vote is summarised in the column headings; for example, 'HousingMinister' relates to a vote on a motion of no confidence in the current Minister for Housing. (See the linked websites above for further details on the 6 votes which took place on November 28th and December 3rd, 4th, 5th, 12th and 18th 2019 respectively).

Interest lies in clustering the TDs to uncover groups with similar attendance and/or voting patterns.

1 (a) Load the voting data into R. Which of the clustering methods that we have seen to date could be used to cluster the TDs based on these binary data? Apply your chosen clustering method to the binary voting data, detailing any decisions you make in the process. How many clusters of TDs do you think are present? [5 marks]

1 (b) Latent class analysis (LCA) can be thought of as a model-based approach to clustering when the recorded data are binary in nature. Polytomous latent class analysis (poLCA) is a clustering method which can be used when variables are categorical. The 2011 *Journal of Statistical Software* paper *poLCA: An R Package for Polytomous Variable Latent Class Analysis* by Linzer and Lewis is posted on Brightspace. Read the Linzer and Lewis (2011) paper to gain some background to and understanding of polytomous LCA. Familiarise yourself with the `poLCA` function in R by reading its help file.

Use the `poLCA` function in R to cluster the TDs based on their voting data. Detail any decisions you make in the process. How many clusters of TDs do you think are present now? On what basis did you make this decision? Include any output or plots which you use to motivate your decision. [10 marks]

1 (c) Compare the clustering from the polytomous analysis 1(b) to the clustering obtained by cutting the dendrogram you view as optimal from (a). [10 marks]

1 (d) Imagine you're a (unbiased!) journalist and your editor asks you to write a report entitled "Voting patterns in the 32nd Dáil Éireann". Examine the membership of the clusters you have found using LCA (the political affiliation of each TD is available on Brightspace in the 'TDs_names_parties.csv' file) and the cluster specific parameters. Write a report outlining your research and conclusions, based on your LCA analysis. Your report should not be longer than 3 A4 pages, in 10 point size font, including background information, some details of the LCA method, plots, tables etc. as you deem necessary. You should include your R code as an Appendix, and this does not contribute towards your page count. Your report should be aimed at a **general scientific audience**. [45 marks]

TOTAL: 70 MARKS

Question 2

2 A sample of 30 wild growing and flowering *Hyptis suaveolens* plants were collected in El Salvador, and the concentrations of 7 terpenes (sabinene, β-pinene, 1.8-cineole, γ-terpinene, fenchone, α-terpinolene and fenchol) were measured (data `Hyptis.csv` available in Brightspace). Interest lies in uncovering different chemotypes of plant (if any) using the terpene measures. The geographical region from which each plant was collected was also recorded, either north, south or east. For the eastern plants, a distinction was made whether the plants grew at low or high altitude.

2 (a) Apply classical metric scaling to the data, explaining any decisions you make in the process. Choose a suitable number of dimensions required to represent the resulting configuration. Explain, using a graphic to support your argument, your reasoning behind your choice of the required number of dimensions. Plot the two dimensional configuration resulting from the application of classical metric scaling. Label each point in your plot using the geographical location of each plant. [10 marks]

2 (b) Apply Sammon's metric least squares scaling and Kruskal's non-metric scaling to the data. Overlay the resulting two dimensional configurations on your plot of the classical scaling configuration. Label each point using the observation number of the associated plant. [5 marks]

2 (c) Use Procrustes analysis to match the three resulting configurations of the plants. Which configurations match best? Suggest a reason for your conclusion. [8 marks]

2 (d) Perform model-based clustering on these data. How many clusters are chosen as optimal? Explain how you arrived at your answer. [7 marks]

TOTAL: 30 MARKS