

Multivariate Analysis

Instructions:

- Use tables, graphs and concise text explanations to support your answers. Unclear answers may not be marked at your own cost. All tables and graphs must be clearly commented and identified.
- You may choose to submit two files, the pdf file of the answers and the R markdown file, containing the R codes, **OR** answer all the questions as an R markdown file. **If you choose to submit pdf file along with the R markdown file, make sure to provide all the answers and interpretations in the pdf file.**
- When answering the questions, just mention the number and part, for example Q1-(a). Avoid copying the whole question in your solutions (increase similarity)

Questions

Question 1. (Multivariate Linear Modeling and Tests of a Covariance Matrix) [40 Marks]

In this question we are going to analyze the class data. This data contains five quiz scores of 53 students, along with information about the gender and major of the students.

- (a) Use `lm()` function to find the relationship between "quiz4" and "quiz5", as response variables, with "quiz1", "quiz2", "quiz3", "gender" and "major" as predictive variables. We call this model **Model 1**. Interpret the result carefully. [10 Marks]
- (b) Now, consider the relationship between (quiz4,quiz5) with (quiz1,quiz2,quiz3). Repeat the same analysis as in Part (a) for these response and predictive variables. We call this model **Model 2**. [10 Marks]
- (c) Compare **Model 1** and **Model 2**. Which one do you prefer? Why? [10 Marks]

- (d) Assume that the random vector $(quiz1, quiz2, quiz3, quiz4, quiz5)^T$ for both genders follow some multivariate normal distributions. Find the covariance matrices of

$$(quiz1, quiz2, quiz3, quiz4, quiz5)^T$$

based on "gender". Can we assume that the covariance matrices for the two groups are the same?? [10 Marks]

Part (e) is optional. You will not miss any point if you do not answer this part. However, if you provide the correct solution you may earn extra points. The Purpose is to show that although some functions may not be introduced in the tutorials, as long as you have the statistical knowledge about the context, you can perform the analysis and provide interpretation of the results.

- (e) To find if there is a significant difference between quiz4 and quiz5 scores between gender and major, use the `manova()` function in R. Interpret the obtained result. Is it in concordance with the result in the previous parts?? [10 Marks]

Question 2. (Factor Analysis) [30 Marks]

The "Harmon23.cor" in the `datasets` package is a correlation matrix of eight physical measurements made on 305 girls between the ages of 7 and 17. You can find the information about this correlation matrix using

`help(Harman23.cor).`

- (a) Perform factor analysis for this correlation matrix using the command

`factanal(factors = m, covmat = Harman23.cor),`

for $m = 1, 2, 3$. Note that using the provided command you do not need to have access to the complete dataset and the factor analysis can be performed using covariance or correlation matrices, with the same reasoning as in PCA. [20 Marks]

- (b) What is the best number of factors in this case?? Interpret the resulting factor loadings for the best case. [10 Marks]

Hint: Not that you cannot rely just on p-values to determine the number of factors in the analysis. You should provide factors that have some interpretation. Therefore, to answer this question, you need to consider factor loadings carefully (Are they large enough??).

Question 3. (Factor Analysis) [30 Marks]

Consider the following sample covariance-covariance matrix:

$$\mathbf{S} = \begin{bmatrix} 1 & 0.4 & 0.9 \\ 0.4 & 1 & 0.7 \\ 0.9 & 0.7 & 1 \end{bmatrix} \quad (0.1)$$

In this question, we are going to follow the steps of Slide 17 in Factor Analysis.

- (a) find the eigenvalues and eigenvectors of \mathbf{S} . [5 Marks]
- (b) Approximate \mathbf{S} using $\lambda_1 \mathbf{e}_1 \mathbf{e}_1^T$ where λ_1 is the largest eigenvalue and \mathbf{e}_1 is the corresponding eigenvector. [5 Marks]
- (c) Estimate the uniquenesses based on the approximation in Part (b). [10 Marks]
- (d) In this Part, we are going to find the exact value of uniquenesses for \mathbf{S} . Write the decomposition of \mathbf{S} , using one factor, as $\mathbf{Q}\mathbf{Q}^T + \mathbf{\Psi}$, where $\mathbf{Q} = (q_1, q_2, q_3)^T$ is the loadings and $\mathbf{\Psi} = \text{diag}(\psi_1, \psi_2, \psi_3)$ are uniquenesses. Find the values of q_i and ψ_i , $i = 1, 2, 3$. Why this solution is improper?? [10 Marks]

Good Luck