Midterm Project

Zhaoxia Yu

2025-04-28

Due on Monday, May 12 2025. R Code should be included as appendices.

In research we often need to compare different methods such as test statistics. Consider three multivariate normal populations. Here I would like you to assess their type I error rate and power of the following four tests for testing $H_0 : \mu_A = \mu_B = \mu_C$ vs $H_1 : H_0$ is not true, where $\mu_A, \mu_B, \mu_C \in \mathbf{R}^p$ are the population mean vectors of the three subpopulations, respectively.

1. Wilk's lambda statistic $\Lambda = \frac{|W|}{|B+W|}$
2. Lawley-Hotelling trace $tr[BW^{-1}]$
3. Pillai trace $tr[B(B+W)^{-1})]$
4. Roy's largest root: the largest eigenvalue of BW^{-1}

Please (1) conduct a simulation study to compare the type I error rate and power of the four tests, (2) apply and compare the methods to a real data set, and (3) write a report to summarize your findings. The real data set should not be a one used in class. It should not be those widely used data sets such as iris, mtcars, airquality, etc. You can use data published in research articles, governments, or institutes.

Part 1: Simulation Study

Data generation. We assume that the three underlying populations have the same variance-covariance matrix Σ. It is helpful to consider different situations. For example, here is a set of four different situations:

1. Independent and equal variance:
$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
2. Positively dependent:
$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$
3. Negatively dependent:
$$\Sigma = \begin{pmatrix} 1 & -0.2 & -0.2 \\ -0.2 & 1 & -0.2 \\ -0.2 & -0.2 & 1 \end{pmatrix}$$

4. Independent but unequal variance:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Use Trial-and-Error to choose reasonable values for $\Sigma, \mu_A, \mu_B, \mu_C$.

Empirical power. To compare the performance of the four methods, you can estimate their power by running simulations. In each simulation, you first generate a random sample from each of the three populations; you then calculate the p-values of the four tests; finally you reject a test if the p-value is less than 0.05. By running B simulations, you can estimate the power of a particular method using the number rejected tests divided by B. B should be chosen such that your estimated power is accurate enough. In my experience, B should be at least a few hundred.

Type I error rate. The method for estimating type I error rate is similar to estimate power empirically, with the difference being that type I error rate is obtained when the null hypothesis is true.

Sample Size. Assume n observations will be obtained from each population. You need to choose n carefully such that the empirical power is neither too small nor too large. This is because when n is too small (large), all methods would have low (high) power and you cannot distinguish their performance. Ideally, you should choose n such that the power of the best method is between 0.8 and 0.9.

Part 2: A Real Data Study

Data. You can use any data set you like. The data set should not be a one used in class. It should not be those widely used data sets such as iris, mtcars, airquality, etc. You can use data published in research articles, governments, or institutes.

Exploratory data analysis and Data Visualization. You should first conduct exploratory data analysis to understand the data. For example, you can use boxplots to visualize the distribution of the data. You can also use appropriate plots to visualize the relationship between different variables.

Analysis. You can use the same four tests as in the simulation study. You can also use the same four situations as in the simulation study.

Part 3: A Report

You are required to write a short report (5 -10 pages, double spaced, including everything except your R code). Your report should cover several sections, such as

1. *Title.* Give a meaningful and informative title to your report.
2. *Abstract.* Provide a short and brief summary to highlight your report. It should include the goal of your study, the methods you use, and your conclusion. Write your abstract in no more than 200 words.
3. *Introduction.* Introduce the purpose of your study.
4. *Simulation Methods.* Describe how you simulate data.

5. *Simulation Results.* Present your results. Please use figures and / or tables to better summarize the results. Including raw R output is not encouraged. R code and output should be included in an appendix.

6. *Real Data.* Describe the data you use and how you analyze it. You can use the same four tests as in the simulation study.

7. *Discussion.* Re-state your main findings and discuss potential pitfalls / limitations of your work. For example, we only look at a few simplified situations for population means and variance-covariance matrices.

8. Appendices. You can include your R code and major output into this section. Additional figures and tables that support the main text can also be included. Data files should be submitted as appendices.

9. References. Optional for this report. If you have read some references of the four tests and want to discuss them, please feel free to include them.