Lisa Fukutoku

May 15, 2023

Statistical Tests Comparison for Multivariate Normal Population

Abstract

Statistical tests play a crucial role in research, enabling us to make informed decisions about population characteristics based on sample data. This study compares four statistical tests in the context of testing a hypothesis regarding the mean vectors of three multivariate normal populations. The tests considered are Wilk's lambda, Lawley-Hotelling trace, Pillai trace, and Roy's largest root. The study is done by examining the type I error rate and power of the four test statistics under three scenarios, including having positively dependent variables, independent but unequal variance, and independent and equal variance. The study employs trial-and-error to select appropriate values for the variance-covariance matrices and population mean vectors. The type I error rate and power are empirically assessed by generating random samples from the populations through simulations, calculating p-values, and determining rejections based on a significance level of 0.05. This study provides valuable insights into the performance of different statistical tests in scenarios involving multivariate normal populations, facilitating researchers in selecting appropriate methods and understanding their limitations.

Introduction

The aim of this study is to conduct a comparative analysis of four statistical tests used in testing a hypothesis regarding the mean vectors of three multivariate normal populations. In this study, we specifically focus on comparing the performance of Wilk's lambda, Lawley-Hotelling trace, Pillai trace, and Roy's largest root tests using their type I error rate and power.

The scenario under investigation involves three trivariate normal populations – a multivariate normal population with three variables – each characterized by a mean vector representing the population means of three subpopulations. The null hypothesis states that the mean vectors of the subpopulations are equal (H_0: $\mu_A = \mu_B = \mu_C$), while the alternative hypothesis accounts for deviations from this equality assumption. We assume that the three populations share the same variance-covariance matrix.

To evaluate the performance of the statistical tests, a simulation approach is employed. We estimate the type I error rate (under the null hypothesis) and power (under the alternative hypothesis) by generating random samples from the multivariate normal populations with simulation and calculating the p-values of the tests. The sample size is carefully selected to ensure sufficient power for distinguishing the tests' performance. This study acknowledges that various situations can arise in practice, such as having positively dependent variables, independent and equal variance, and independent but unequal variance. By exploring these situations and evaluating the tests' performance within each, we aim to provide useful insights into the strengths and limitations of the tests.

In this paper, we now will estimate and evaluate the type I error rate and power of the four statistical tests under different scenarios, which will enhance the reliability and accuracy of statistical analyses in multivariate data settings.

Simulation Methods

To evaluate the performance of the four statistical tests, a simulation modeling was conducted for the estimation of the type I error and power of the tests. We first defined three scenarios for comparing the effectiveness of each test:

1. Positively dependent variables: two populations have same mean vectors of (0,0,0) and one has different mean (2,2,2).

	[,1]	[,2]	[,3]
[1,]	1.0	0.8	0.6
[2,]	0.8	1.0	0.4
[3,]	0.6	0.4	1.0

Table 1: variance-covariance matrix Σ for Scenario 1

2. Variables with independent but unequal variance: for the comparison purpose, we set the same as scenario 1, two populations have same mean vectors of (0,0,0) and one has different mean (2,2,2).

	[,1]	[,2]	[,3]
[1,]	0.5	0.0	0.0
[2,]	0.0	1.5	0.0
[3,]	0.0	0.0	2.5

Table 2: variance-covariance matrix Σ for Scenario 2

3. Variables with independent and equal variance: two populations have same mean vectors of (0,0,0) and one has different mean (-3,-3,-3) with the simplest variance-covariance matrix.

	[,1]	[,2]	[,3]
[1,]	1	0	0
[2,]	0	1	0
[3,]	0	0	1

Table 3: variance-covariance matrix Σ for Scenario 3

The simulation methods involved several steps. In each simulation, we first generated a random sample from each of the three multivariate normal populations under consideration. The sample size, denoted as “n”, was carefully chosen such that the empirical power is neither too small not too large. The selection of the sample size aimed to achieve a power range between 0.8 and 0.9 to distinguish the performance of the tests. The generated sample data reflected the assumed equal variance-covariance matrix and mean vectors of the respective subpopulations.

Secondly, the four statistical tests were applied for each simulated sample to test the null hypothesis of equal mean vectors (H_0: $\mu_A = \mu_B = \mu_C$) against the alternative hypothesis of unequal mean vectors (H_a: the null is not true). The test statistics of interest were Wilk's lambda, Lawley-Hotelling trace, Pillai trace, and Roy's largest root. The test statistics were calculated based on the sample means and the estimated variance-covariance matrix.

Thirdly, from the obtained test statistics, p-values were computed for each test. The p-values indicated the probability of observing test statistics equal to or more extreme than the ones observed, assuming the null hypothesis was true. A significance level of 0.05 was chosen as the threshold for hypothesis rejection. If the calculated p-value for a specific test was less than 0.05, the null hypothesis was rejected, indicating evidence of a significant difference among the mean vectors.

Lastly, to estimate and evaluate the type I error and power of each test, the simulation was replicated for 300 times. The type I error rate of a test was obtained by calculating the proportion of replications in which the null hypothesis was erroneously rejected, while the power was determined by computing the proportion of replications in which the test rejected the null hypothesis correctly. The R programming language was utilized to implement the simulation methods and perform the necessary calculations.

Simulation Results

From the analysis, the null hypothesis was rejected for all four tests in all three situations at a significance level of 0.05. This implies that there is evidence to suggest that the mean vectors of the three multivariate normal populations are not equal. As we set one population to have a different mean vector for all three scenarios, this result is what we expected.

The type I error rate refers to the probability of incorrectly rejecting the null hypothesis when it is actually true. Figure 1-3 demonstrate that the type I error rates were generally well-controlled at a significance level of 0.05 across different scenarios for all tests except Roy's largest root test, which had a remarkably large error rate. Specifically, there was a small tendency shown in the first and second scenario that Wilk's lambda test achieved the smallest error rate, followed in order by Lawley-Hotelling trace test and Pillai trace test.

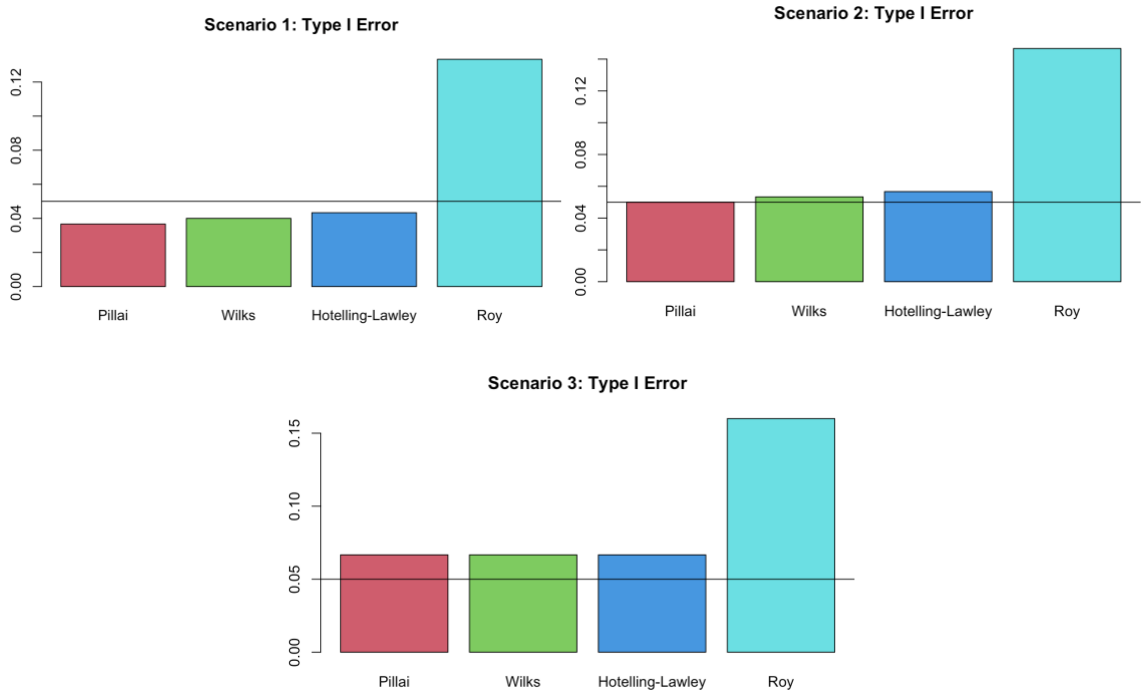


Figure 1-3: Type I error rate of four test for scenario 1, 2, and 3

The power refers to the probability of correctly rejecting the null hypothesis. Figure 4-6 represent the power of four tests for three situations. Figure 5 and 6 indicate that all tests, with the exception of Roy's test, achieved satisfactory power in a rage between 0.8-0.9 in detecting significant differences among the mean vectors when the null hypothesis was false. However, the overall power levels were lower in the first scenario when the variables are positively dependent.

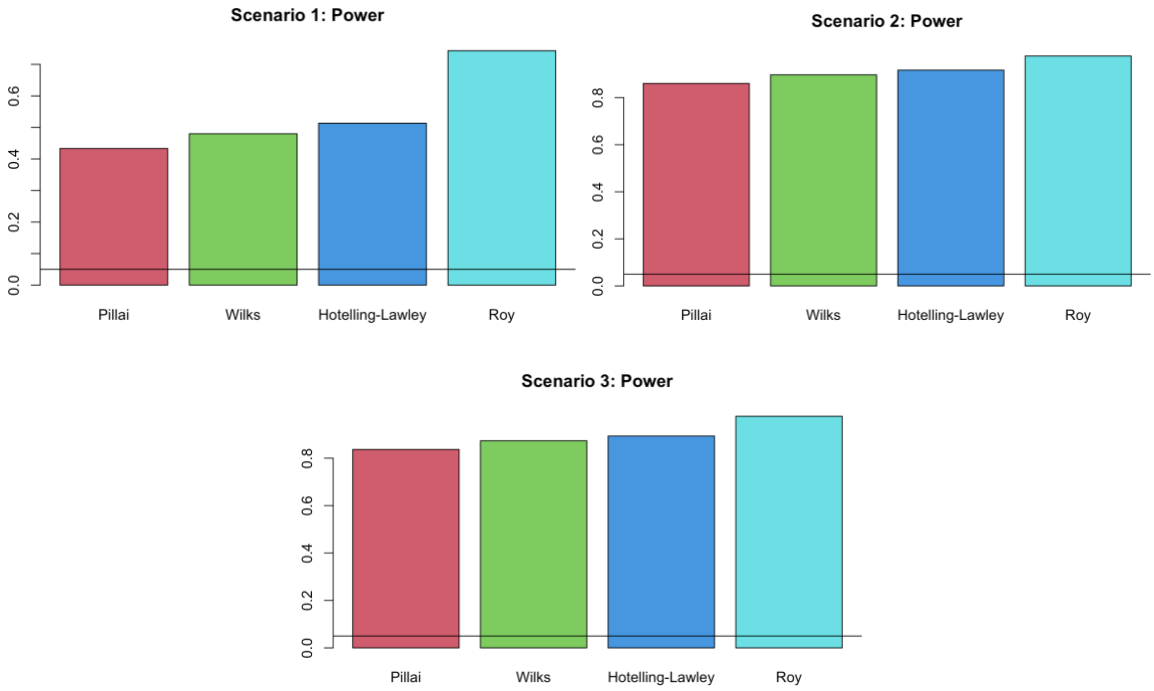


Figure 4-6: Power of four test for scenario 1, 2, and 3

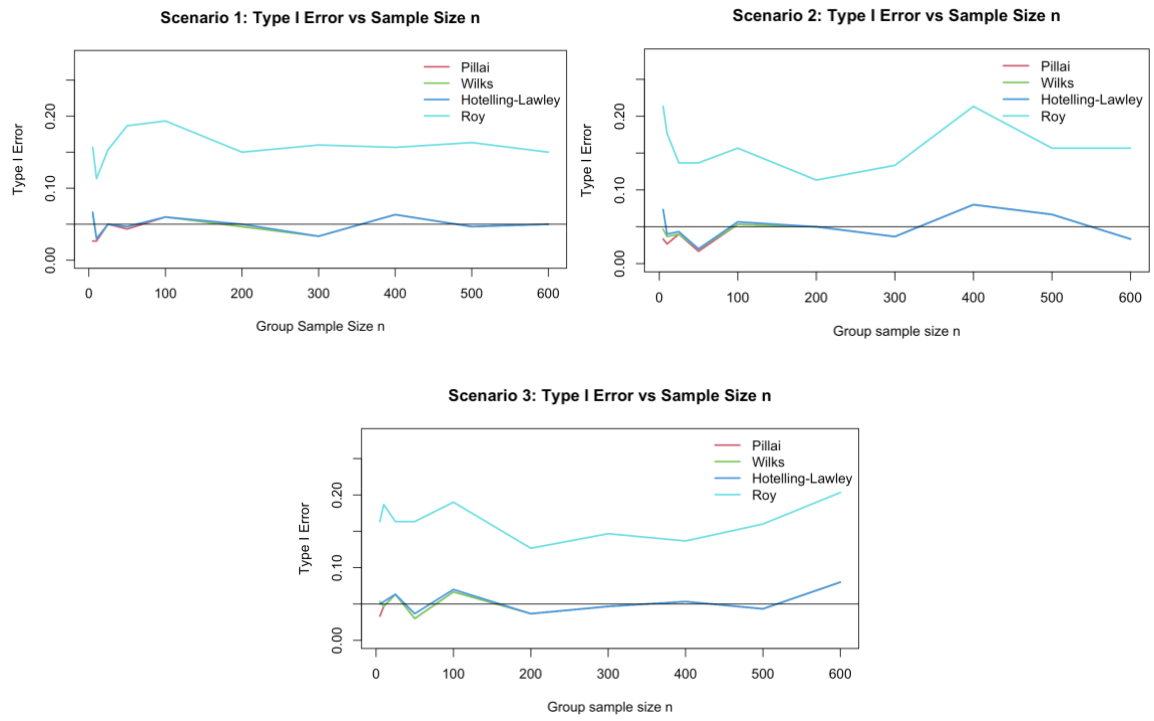


Figure 7-9: Type I error rate vs Sample size "n" of four test for scenario 1, 2, and 3; indicating how type I error rate changes while increasing the sample size.

Figure 7-9 illustrate how type I error rate changes while increasing the sample size. While the Roy's test shows significantly larger error rates than other tests, the three tests are all relatively consistent around 0.05 type I error rate even when the sample size changes from 0 to 600.

Furthermore, if we assume that two populations share the same mean vector of (0,0,0), while the third population has a mean vector of (c,c,c), we can examine the power as a function of c. In other words, we can consider power(c) as the probability of rejecting the null hypothesis when one group has a mean vector of (c,c,c) instead of (0,0,0). Notably, when c is equal to 0, the type I error is recovered. The following plot figures (Figure 10-12) are the comparison of the power function in different settings – (c,c,c) versus (c1,0,0) versus (0,c2,0) versus (0,0,c3) – depicting how power changes when c, c1, c2, and c3 increase. Although there are no significant differences in plots of c1, c2, c3 across three scenarios, the plots of c values appear to have higher power with smaller c values than values for c1, c2, and c3. This is because the value c was applied for all three mean values in a vector as opposed to c1, c2, and c3, which were only employed for a single value in a vector surrounded by 0.

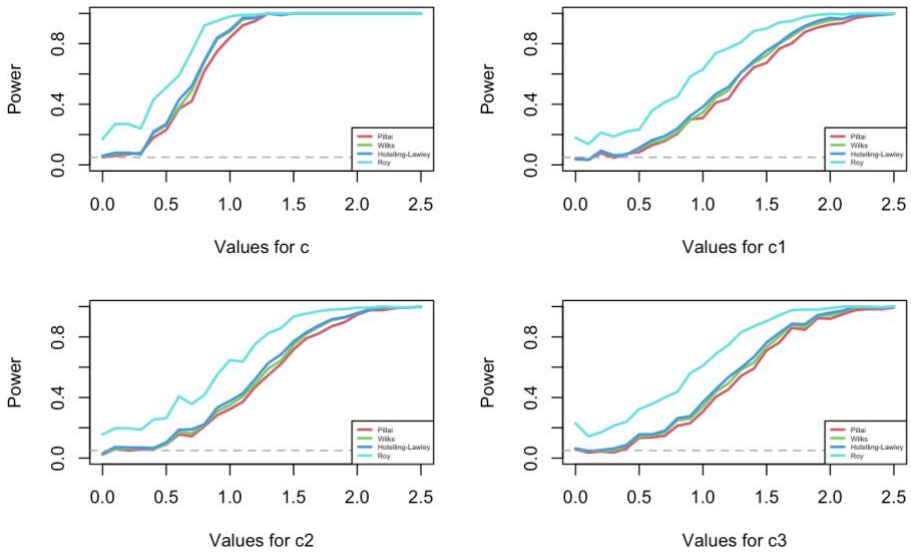


Figure 10: Scenario 1: Power Function (c,c,c) vs (c1,0,0) vs (0,c2,0) vs (0,0,c3)

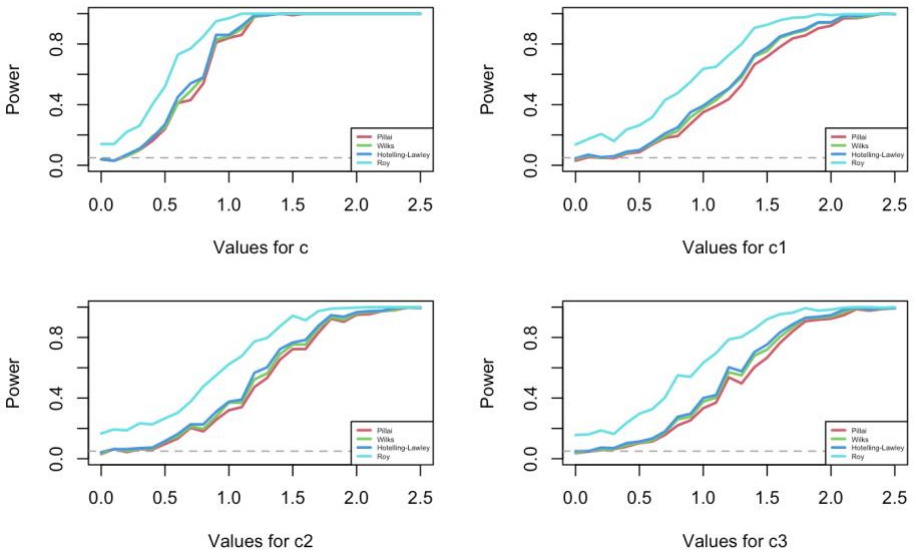


Figure 11: Scenario 2: Power Function (c,c,c) vs (c1,0,0) vs (0,c2,0) vs (0,0,c3)

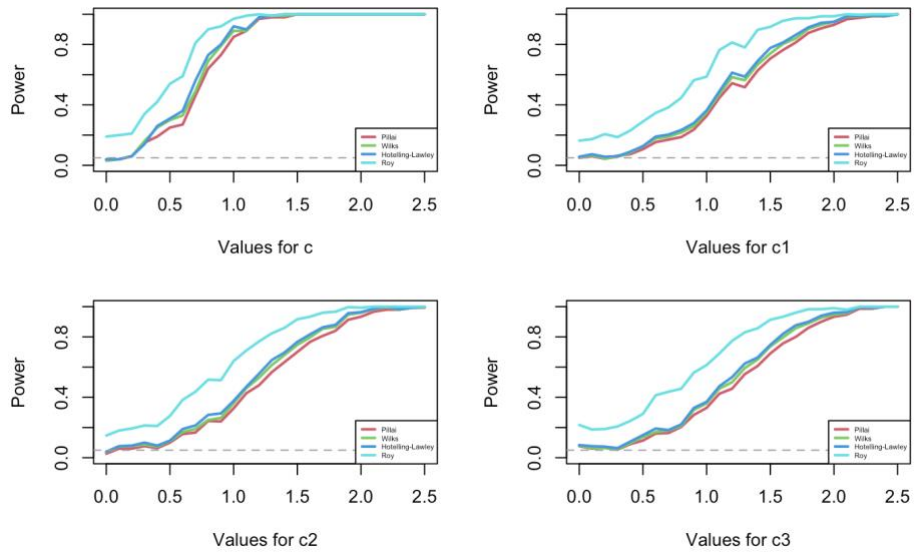


Figure 12: Scenario 3: Power Function (c,c,c) vs (c1,0,0) vs (0,c2,0) vs (0,0,c3)

Overall, the simulation results highlight the characteristics of the four statistical tests' performance in terms of type I error rate and power. In particular, the analysis consistently revealed poor results for Roy's largest root test across all scenarios.

Discussion

The simulation results shed light on the performance characteristics of the four statistical tests – Wilk's lambda, Lawley-Hotelling trace, Pillai trace, and Roy's largest root – for comparing mean vectors of multivariate normal populations. The type I error rate estimates demonstrated that all tests, except Roy's largest root, maintained their nominal level of significance at the selected threshold of 0.05 across the different scenarios. This indicates that the three tests provide reliable inference and control the probability of incorrectly rejecting the null hypothesis when it is true. The well-controlled type I error rate ensures the validity of the reported significant differences among the mean vectors.

The results of power estimation also shared the similar trait from type I error estimation results that three tests, Wilk's lambda, Lawley-Hotelling trace, and Pillai trace, demonstrated relatively close levels of power across all conditions. These tests achieved power in an adequate rage between 0.8-0.9, whereas the overall power levels dropped in the first scenario. The tests that indicated higher power in detecting significant differences among the mean vectors are more suitable, as they have a higher probability of identifying true effects in the population. These findings enhance our understanding of the strengths and limitations of each test and emphasize the need for researchers to carefully assess the characteristics of their data and make informed decisions in selecting the most suitable test for their analysis.

Additionally, it is important to acknowledge certain limitations of the study. In this study, the simulations were conducted under simplified scenarios, assuming equal variance-covariance matrices for all populations and specific mean vectors; however, real-world data may have more complex distributions that could impact the performance of the tests.

In conclusion, this study provides valuable insights into the four statistical tests for comparing mean vectors of multivariate normal populations by evaluating their type I error rate and power. It is essential to consider the specific characteristics of the data and research context when selecting an appropriate test for hypothesis testing. Considering the Roy's largest root test consistently demonstrated poor results through the analyses, we suggest that examining these tests without Roy's test would enable us to clearly identify the differences in three tests, Wilk's lambda, Lawley-Hotelling trace, and Pillai trace.

Appendix

Data generation:

Note: for the scenario 2, sigma values were changed; Sigma = diag(c(0.5, 1.5, 2.5)).

For the scenario 3, mu2 <- c(-3, -3, -3) and Sigma = diag(1, 3).

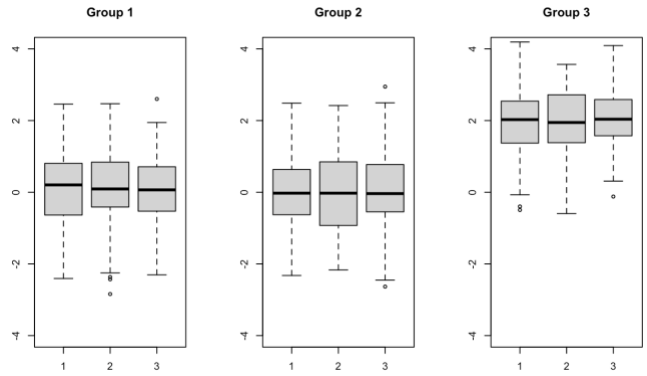
```
# Scenario 1: Sigma = Positively dependent
n <- 100  # Number of samples
G <- 3   # Number of population groups
mu2 <- c(2, 2, 2)  # Mean for group 2 (3D)
p <- length(mu2)   # Number of dimensions, 3D

# Data generation function
generate_data <- function(n, G, mu2, mu1 = rep(0, p), Sigma = matrix(c(1, 0.8, 0.6, 0.8, 1, 0.4, 0.6, 0.4, 1), nrow = 3, ncol = 3)){
  Y <- matrix(nrow = 0, ncol = p)  # Initialize an empty matrix

  for (g in 1:ceiling(G/2)){
    Y <- rbind(Y, MASS::mvrnorm(n, mu1, Sigma))  # Generate samples from group 1
  }
  for (g in 1:floor(G/2)){
    Y <- rbind(Y, MASS::mvrnorm(n, mu2, Sigma))  # Generate samples from group 2
  }
  Y
}

Y <- generate_data(n, G, mu2)
```

Data for the scenario 1:




```
# Hypothesis testing function
# 1 = reject; 0 = accept

test <- function(n, G, Y){
  groups <- rep(c(paste("Group", 1:G)), each=n)
  obj <- manova(Y ~ groups)
  tests <- c("Pillai", "Wilks", "Hotelling-Lawley", "Roy")
  reject <- rep(0, 4)

  for (t in 1:length(tests)){
    reject[t] <- summary(obj, test = tests[t])$stats[1,6]<0.05
  }
  reject
}

results <- test(n, G, Y)
names(results) <- c("Pillai", "Wilks", "Hotelling-Lawley", "Roy")
results
```

```
          Pillai            Wilks Hotelling-Lawley              Roy
              1                1                1                1
```

```r
# Simulation function
simulate <- function(B, n, G, mu2, Sigma = diag(1, 3)){
  results <- rep(0,4)
  for (b in 1:B){
    Y <- generate_data(n, G, mu2, Sigma = Sigma)
    results <- results + test(n, G, Y)
  }
  results/B
}
```

```r
# Type I error plot
tests <- c("Pillai", "Wilks", "Hotelling-Lawley", "Roy")
alpha <- simulate(B = 300, n = 100, G = 3, mu2=c(0,0,0), Sigma = matrix(c(1, 0.8, 0.6, 0.8, 1, 0.4, 0.6, 0.4, 1), nrow = 3, ncol = 3))
names(alpha) <- tests
barplot(alpha, col=c(2:5))
abline(h=0.05)
title('Scenario 1: Type I Error')
```

```r
# Type I Error vs Sample Size n plot
N <- c(5, 10, 25, 50, 100, 200, 300, 400, 500, 600)
alpha <- matrix(0, length(N), 4)

# Positively dependent variance for Sigma (variance-covariance matrix) for all 3 groups
for (k in N){
  alpha[which(k==N), ] <- simulate(B = 300, n = k, G = 3, mu2=c(0,0,0),
                                   Sigma = matrix(c(1, 0.8, 0.6, 0.8, 1, 0.4, 0.6, 0.4, 1), nrow = 3, ncol = 3))
}
plot(N,rep(0, length(N)),type="n", ylim=c(0, .28), xlab="Group Sample Size n", ylab="Type I Error")

for (i in 1:4){
  lines(N, alpha[, i], col=i+1, type = 'l', lwd=2)
}
abline(h=0.05)
legend("topright", legend = tests, col = c(2:5), lwd = rep(2, 4), bty = "n")
title('Scenario 1: Type I Error vs Sample Size n')
```

```r
# Power plot
tests <- c("Pillai", "Wilks", "Hotelling-Lawley", "Roy")
alpha <- simulate(B = 300, n = 10, G = 3, mu2=c(1,1,1), Sigma = matrix(c(1, 0.8, 0.6, 0.8, 1, 0.4, 0.6, 0.4, 1), nrow = 3, ncol = 3))
names(alpha) <- tests
barplot(alpha, col=c(2:5))
abline(h=0.05)
title('Scenario 1: Power')
```

```r
# Power Function (c,c,c) vs (c1,0,0) vs (0,c2,0) vs (0,0,c3) plots

# Run the simulation for each value of C
C <- seq(0, 2.5, .1)

# Initialize matrices to store the results
results <- matrix(0, length(C), 4)
results1 <- matrix(0, length(C), 4)
results2 <- matrix(0, length(C), 4)
results3 <- matrix(0, length(C), 4)

for (i in 1:length(C)){
  results[i, ] <- simulate(100, 10, 3, c(C[i],C[i],C[i]))
  results1[i, ] <- simulate(B = 300, n = 10, G = 3, mu2 = c(C[i], 0, 0))
  results2[i, ] <- simulate(B = 300, n = 10, G = 3, mu2 = c(0, C[i], 0))
  results3[i, ] <- simulate(B = 300, n = 10, G = 3, mu2 = c(0, 0, C[i]))
}

# Set up the plot layout
par(mfrow=c(2, 2))

# Adjust margins (bottom, left, top, right)
par(mar=c(4, 4, 2, 2))

# Plot the power for each test as a function of C
plot(0,0,type="n", xlim=c(min(C), max(C)), ylim=c(0, 1),
     xlab="Values for c", ylab="Power")
for (i in 1:4){
  lines(C, results[, i], type = "l", col = i+1, lwd=2)
}
abline(h=0.05, col = 8, lty=2)
legend("bottomright", legend = tests, col = c(2:5), lwd = rep(2, 4), cex=0.4)
```

```r
# Plot the power for each test as a function of C for the first dimension
plot(0, 0, type="n", xlim=c(min(C), max(C)), ylim=c(0, 1),
     xlab="Values for c1", ylab="Power")
for (i in 1:4){
  lines(C, results1[, i], type = "l", col = i+1, lwd=2)
}
abline(h=0.05, col = 8, lty=2)
legend("bottomright", legend = tests, col = c(2:5),
       lwd = rep(2, 4), cex=0.4)

# Plot the power for each test as a function of C for the second dimension
plot(0, 0, type="n", xlim=c(min(C), max(C)), ylim=c(0, 1),
     xlab="Values for c2", ylab="Power")
for (i in 1:4){
  lines(C, results2[, i], type = "l", col = i+1, lwd=2)
}
abline(h=0.05, col = 8, lty=2)
legend("bottomright", legend = tests, col = c(2:5),
       lwd = rep(2, 4), cex=0.4)

# Plot the power for each test as a function of C for the third dimension
plot(0, 0, type="n", xlim=c(min(C), max(C)), ylim=c(0, 1),
     xlab="Values for c3", ylab="Power")
for (i in 1:4){
  lines(C, results3[, i], type = "l", col = i+1, lwd=2)
}
abline(h=0.05, col = 8, lty=2)
legend("bottomright", legend = tests, col = c(2:5),
       lwd = rep(2, 4), cex=0.4)
```