Two-sample Testing

I-Lun, Tu

Introduction

Discerning the differences between two samples is a classic and important issue of statistical testing. In business settings, this statistical procedure is often applied to understand whether two groups of potential customers behave similarly, such as A/B testings. With this information, marketing managers can decide why and how to tailor different marketing campaigns to better cater to potential customers. However, it is often not obvious which kinds of tests one should use when confronted with various different constrains posed by the data at hand, such as non-normal distributions, high dimensions, small sample size, etc. Therefore, in this report, we try to address these issues and examine various tests to see how they perform.

We consider 4 different test statistics and we use permutation test to obtain the significance level of the resulting statistics. In this way, we do not have to make use of further approximation methods when the assumed underlying distributions for the statistics in consideration are not attained. For each test, Type 1 and Type 2 error are the average of 100 decisions and each decision is the result of 500 times of permutation tests.

- Nearest Neighbor Test (NNT)
- Energy-based Test (EBT)
- Hotelling's T-squared Test (HTT)
- Graphed-based Two-sample Test (GST)

The efficacy of a statistical test depends on many factors. Listed below are the factors considered in this report. After understanding the advantages and disadvantages of the proposed tests, we apply them in the examination of the prostate cancer dataset. We consider 4 variables and are interested in understanding if there is an overall difference between people older than 65 and younger than 65.

- Sample size: how many samples we need for the test to be effective.
- Dimension: how well or how bad a test performs in low and high dimensions.
- Distribution: how sensitive a test responds to data generated from different distributions.
- Efficiency: how much computational power a test requires.

We discover that EBT and HTT tests are more powerful tests than the others in this particular setting. Therefore, We came to the conclusion that there is indeed an overall difference despite the fact that NNT and GST point in different direction.

Analysis

1. Test Tuning

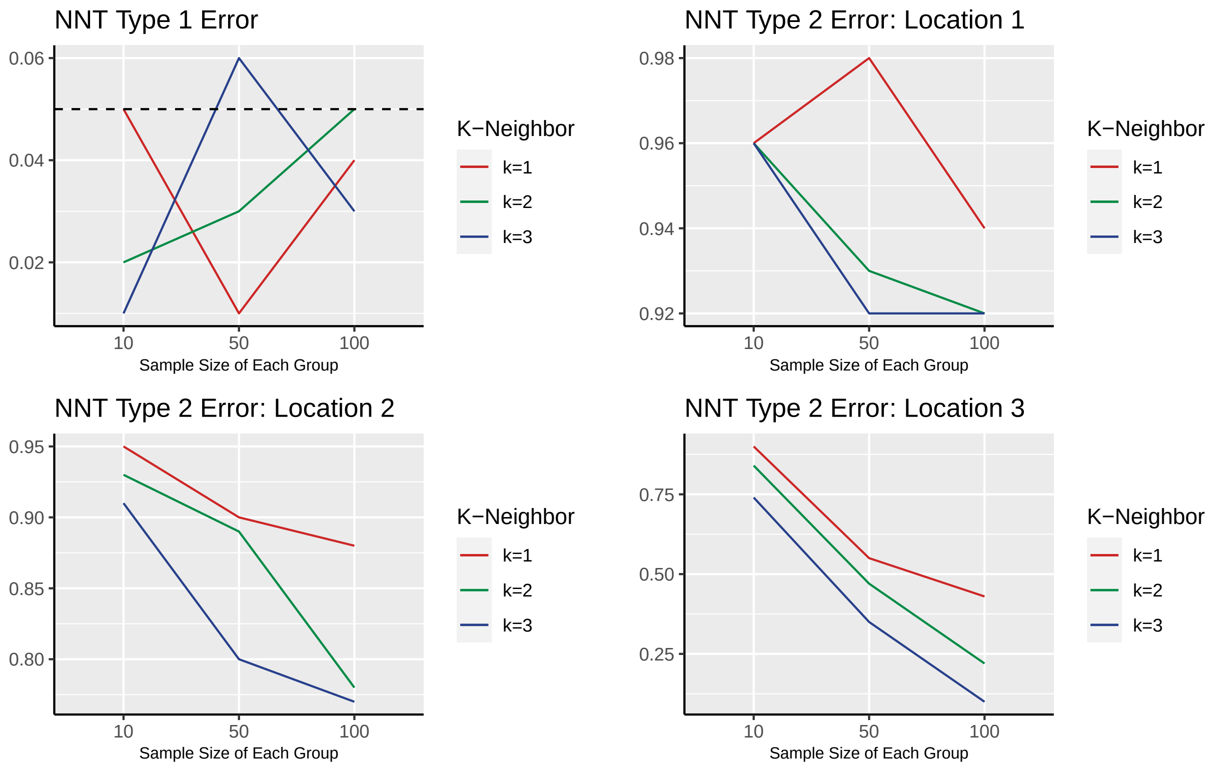
Before comparing directly each tests for effectiveness, we note that NNT, EBT, and GST tests are open to customization, meaning that we can adjust some variables within the tests to better detect differences in data. Therefore, we start from adjusting these tests first.

(1). For the following results, we use the same size of samples for each group to simulate what we will encounter for testing differences:

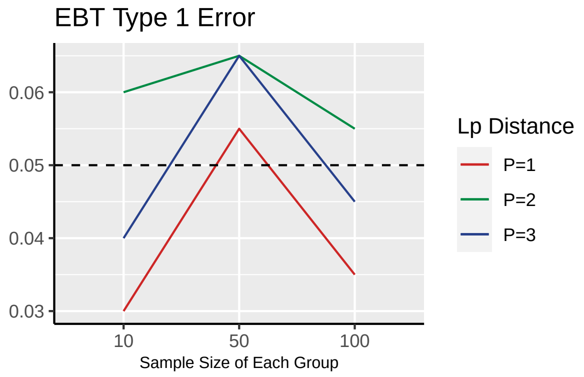
- First, we see whether we can control the Type 1 error, which is the risk that we take on when we make the decision that the effect of two treatments are not similar when in fact they are identical.
- Secondly, in contrast, after controlling the Type 1 error, we observe how effective the tests are in detecting differences in treatment effects when in fact they are different. In this scenario, another risk, Type 2 error, incurs, which are the risk that we would fail to report the differences. Note we provide three cases to test these risks. Ranging from hardest to easiest to detect are in order 1, 2, 3.

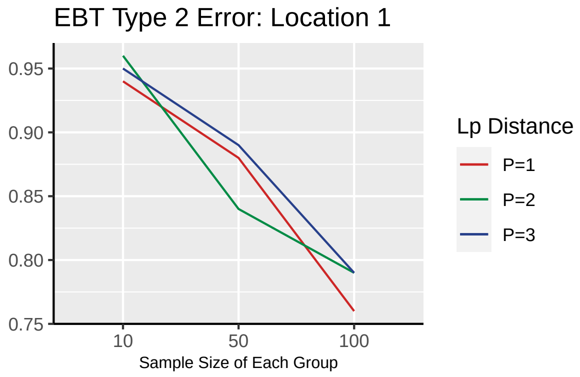
(2). From the below graph, we successfully control the Type 1 error for each test under the dashed line when sample size at each group reaches 100. For GST, We report the best combination here. We conclude that the best customizations are NNT-K=3, EBT-P=2, and GST-P=2 & Q=3.

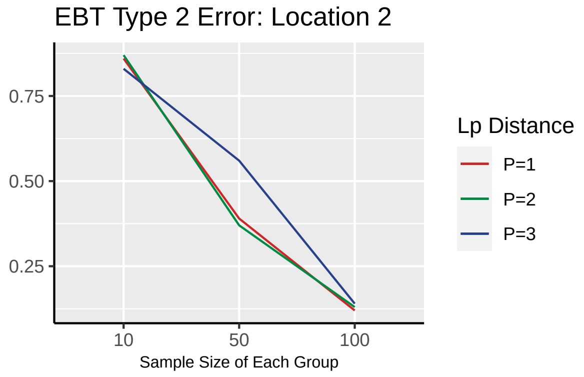
- NNT

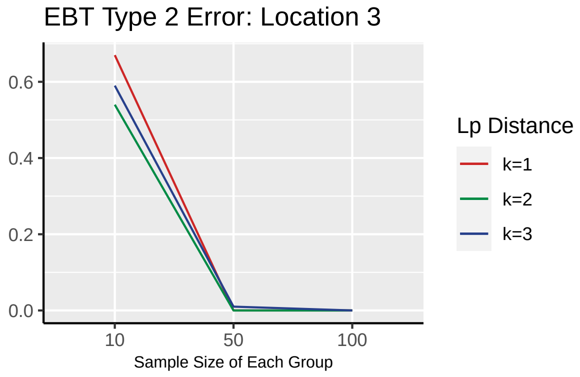


- EBT

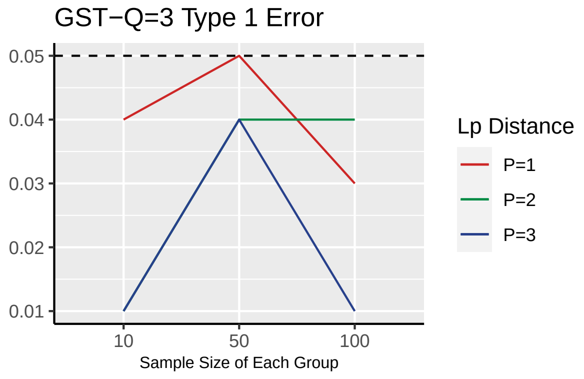

EBT Type 1 Error

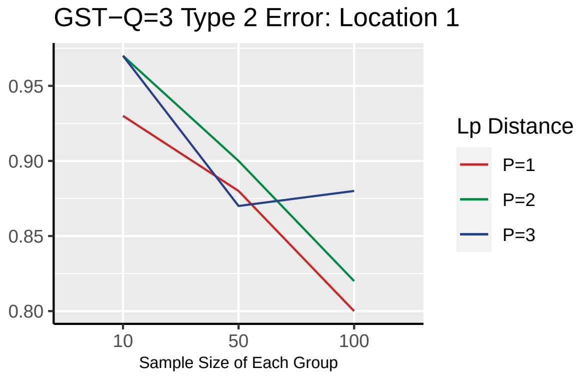

EBT Type 2 Error: Location 1

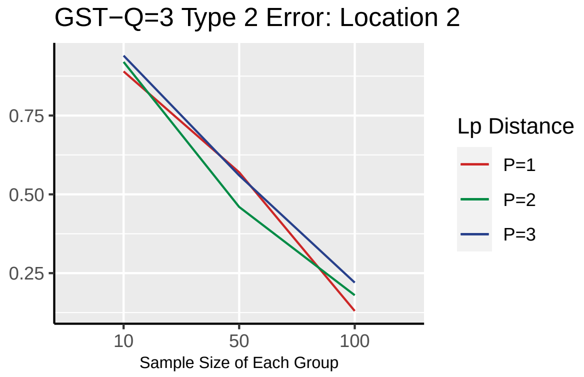

EBT Type 2 Error: Location 2

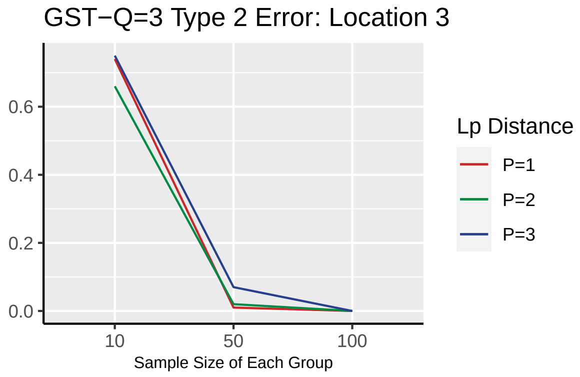

EBT Type 2 Error: Location 3

- GST


GST−Q=3 Type 1 Error


GST−Q=3 Type 2 Error: Location 1


GST−Q=3 Type 2 Error: Location 2


GST−Q=3 Type 2 Error: Location 3

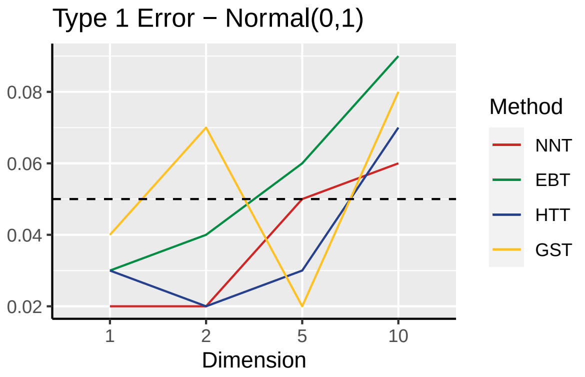
2. Test Comparison

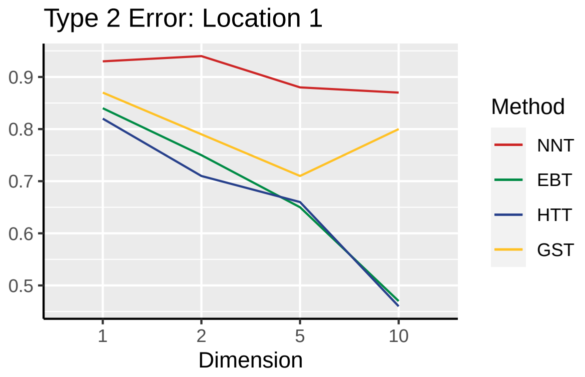
(1) Dimension: we use 100 samples for each group. We discover that Type 1 error cannot be controlled properly when dimension are higher than 5 for most of the tests. For type 2 error, EBT and HTT are effective for all dimension, while NNT and GST are only suitable for high dimension.

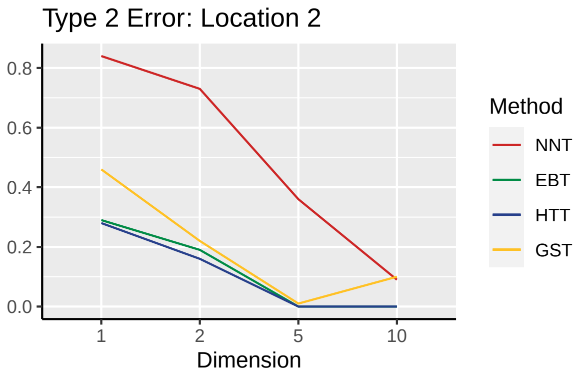
(2) Sensitivity: we use different types of data to compare (Gamma), and discover that Type 1 error can be controlled except for GST. For Type 2 error, we find out that EBT and HTT are consistent, while NNT and GST are not effective even when the differences are obvious.

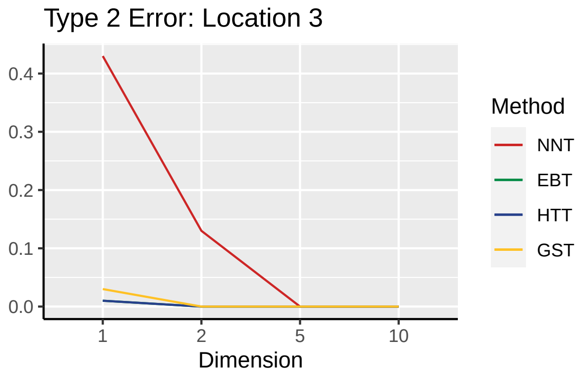
(3) Efficiency: The lower the amount of time required, the more efficient the test is. We compare the execution time for running ten times of each test for dimension 2, 5, and 10. We find out that EBT and HTT are more efficient than the others, while NNT is computationally costly especially when dimension is high.

- Dimension

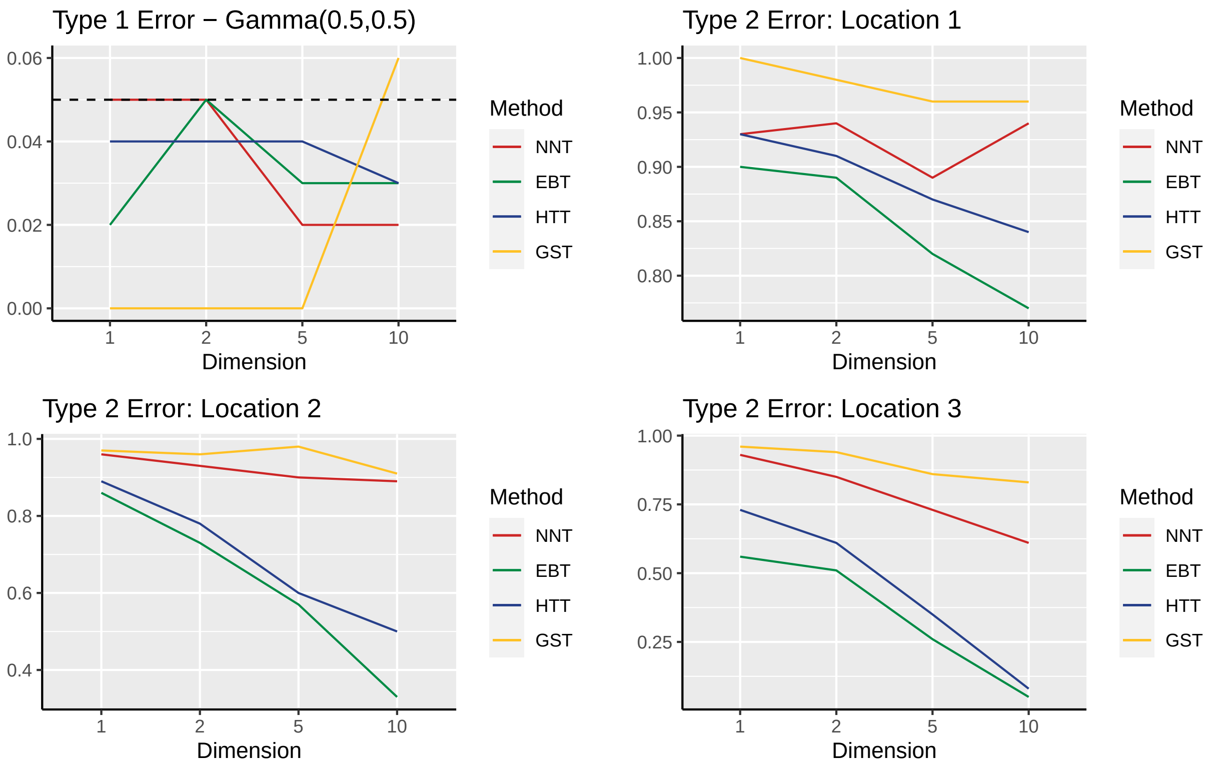

Type 1 Error − Normal(0,1)


Type 2 Error: Location 1

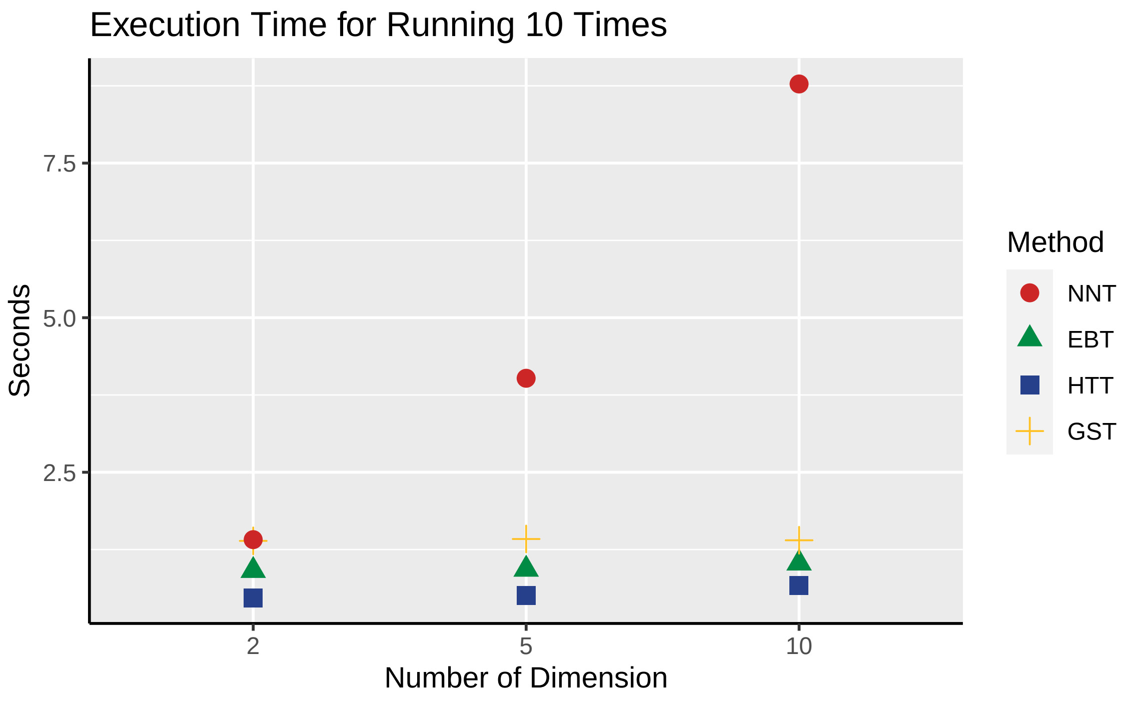

Type 2 Error: Location 2


Type 2 Error: Location 3

- Distribution/ Sensitivity



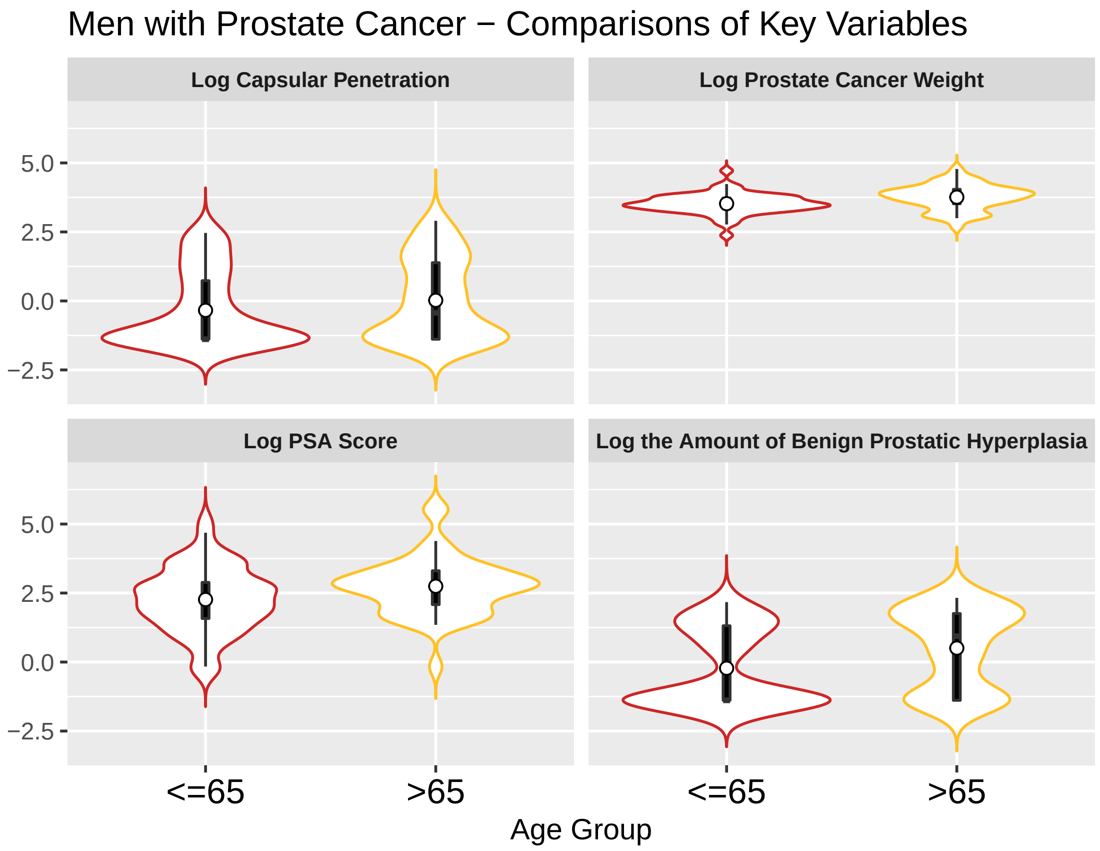
- Efficiency



Application to Prostate Cancer Dataset

In this application, we consider 4 variables (4 dimension). From the below density approximation plots, we see that means, the white dot, are quite close in value and also the general appearance of the density curves are similar. To understand the quantitative differences, we must perform statistical tests. With alpha level set at 0.05, we conclude the following:

- EBT and HTT conclude that there is indeed an overall difference, while GST and NNT conclude the contrary.
- In this data set, we have around 50 samples for each group and we test only for 4 dimension. We know that HTT and EBT performs better in these settings.
- We conclude that there is indeed an overall difference between people younger and older than 65, since EBT and HTT are more powerful and reliable tests in this case.



Conclusion and Discussion

Comparing the tests in various criteria, we suggest EBT test for its overall reliability across many data structures, effectiveness in controlling Type 1 error and lowering Type 2 error, and efficiency in execution. However, we note that to produce consistent results in testings, sample size for each group should at least be above 50 and dimension should be lower than 5.