Exploiting the Classification Algorithm for Robust Multidimensional Homogeneity Test in Causal Inference

MA 590 Special Topics: Causal Inference

Aukkawut Ammartayakun

Worcester Polytechinic Institute

19 March, 2023

Introduction

Motivation

- One of the assumption in causal inference is the **ignorability** assumption.
    - The ignorability assumption is violated when the treatment assignment is not random.
- Hypothesis testing in traditional manner does not work well with multidimensional data like image or text data.
    - This problem can be viewed as classification problem (whether to reject the null hypothesis or not).
    - (Hopefully) this also leads to explainability.

Problem Statement and Possible Solution

Let $f : \mathcal{D} \to [0, 1]$ (classification algorithm) such that $f(x, y)$ for $(x, y) \in \mathcal{X} \supseteq \mathcal{D}$ reflecting the probability of rejecting the null hypothesis $H_0 : X = Y$ for paired data (x, y) in the data space $\mathcal{X}$.

- p-value can be estimated from f and the decision can be done with thresholding.

Problem: How can we search for such f? What would be the power of the test?

Previous Works

Previous Works

Let say we have two group of data D: D_t and D_c. We want to test whether there is a difference between the two groups.

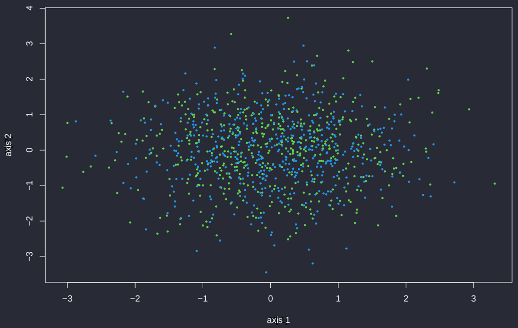
Classifier Two-Sample Test (Lopez-Paz and Oquab 2017)

- Combined two dataset into one dataset D.
- Split the dataset into training and testing set.
- Fit the classifier (like logistic regression) to the training set and predict the testing set.
- Calculate the empirical loss l_e of the classifier. If $|l_e - 0.5| < \epsilon$, then $\bar{\tau} = 0$.

Testing the Algorithm

```r
set.seed(590)
# generate random multivariate gaussian data
n <- 1000
d <- 2
X <- matrix(rnorm(n*d), n, d)
y <- c(rep(0, n/2), rep(1, n/2))
c2st(X, y, echo = TRUE)
```

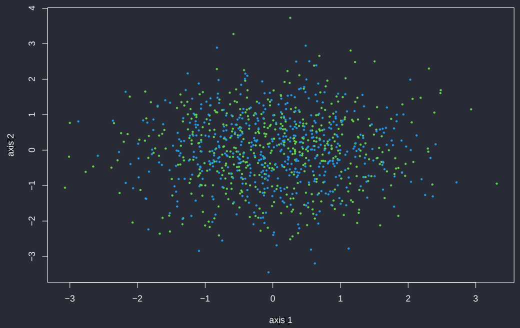
```
[1] Empirical loss: 0.57
[1] p-value: 0.282762829938323
```



Testing the Algorithm

```r
set.seed(590)
# generate random multivariate gaussian data
n <- 1000
d <- 2
X <- matrix(rnorm(n*d), n, d)
y <- c(rep(0, n/2), rep(1, n/2))
c2st(X, y, echo = TRUE)
```

```
[1] Empirical loss: 0.57
[1] p-value: 0.282762829938323
```

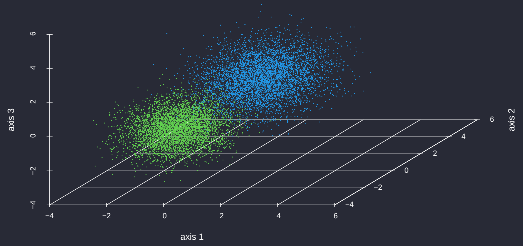


It is the *impossible* classification problem. Thus, the result should be close to near-chance level.

Testing the Algorithm

```r
set.seed(590)
# generate two dataset: two gaussians
n <- 5000
d <- 3
X0 <- matrix(rnorm(n*d, -1,0.8), n, d)
X1 <- matrix(rnorm(n*d, 1,1), n, d)
y <- c(rep(0, n), rep(1, n))
c2st(rbind(X0, X1), y, echo = TRUE)
```

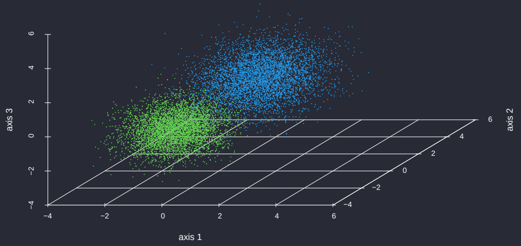
```
[1] Empirical loss: 0.0285
[1] p-value: 0
```



Testing the Algorithm

```r
set.seed(590)
# generate two dataset: two gaussians
n <- 5000
d <- 3
X0 <- matrix(rnorm(n*d, -1,0.8), n, d)
X1 <- matrix(rnorm(n*d, 1,1), n, d)
y <- c(rep(0, n), rep(1, n))
c2st(rbind(X0, X1), y, echo = TRUE)
```

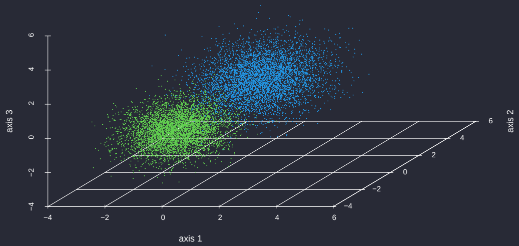
```
[1] Empirical loss: 0.0285
[1] p-value: 0
```



Test statistic diverges from near-chance level as there is an exist of "linear" decision boundary. The dimensionality problem in homogenity test is solved.

Question: Does changing from GLM to other models increases (widen the range of) the testing power, β?

Testing the Algorithm



Comparison to Maximum Mean Discrepancy

Use C2ST on Causal Inference

- But, how can we use C2ST on causal inference?

Example:

Proposed Solution

Possible Solution

- Combine both treatment and control group within each stratum into one dataset D_i^s.
- For each group, fit the classifier (like logistic regression) to the training set and predict the testing set.
- Calculate the empirical loss l_e of the classifier. If $|l_e - 0.5| < \epsilon$, then $\bar{\tau}_{\text{within}} = 0$
- Find the way to infers $\bar{\tau}_{\text{between}}$

Results

Results (Con't)

Conclusion

References

References

Lopez-Paz, David, and Maxime Oquab. 2017. “Revisiting Classifier Two-Sample Tests.” In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJkXfE5xx.