Introduction to Targeted Maximum Likelihood Estimation Likelihood

Xiqian Yuan

1 Introduction

In the realm of machine learning, achieving robust and reliable inference is a paramount goal. However, the complexity of real-world scenarios often renders traditional assumptions unattainable. Enter Targeted Maximum Likelihood Estimation (TMLE), a powerful methodology that embraces the challenge of making few assumptions while providing double robustness in its inferences. By seamlessly integrating causal inference principles, TMLE stands out as a versatile tool capable of handling intricate data landscapes. Its ability to thrive in scenarios with minimal assumptions makes it an invaluable asset for researchers and practitioners seeking dependable insights in the face of uncertainty, making TMLE a compelling avenue for advancing the frontiers of robust machine learning.

2 Algorithm

2.1 Data Structure

First, I will show the data structure needed to define in our project. We now only consider a scenario with a binary outcome and treatment, aiming to estimate the Average Treatment Effect (ATE) under causal assumptions. The data structure encompasses variables such as the binary outcome Y representing the treatment response. the treatment/exposure variable A, and confounders W.

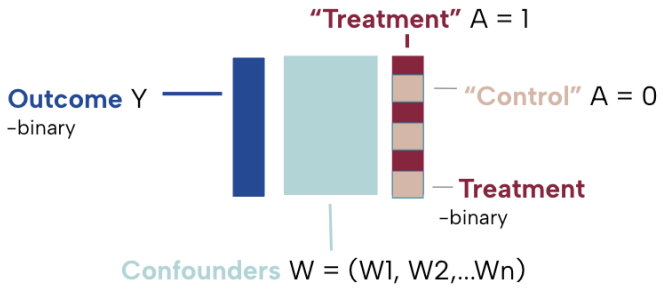


Figure 1: Data Structure used in TMLE

2.2 Initial Outcome

To initiate our TMLE process, we fit machine learning model like lasso, random forest, etc to estimate mean functions using the treatment and confounders to predict outcome. Also, expanding from the model, we then calculate two mean effect under each treatment and control. Then we get our two mean effects which are our two initial outcomes.

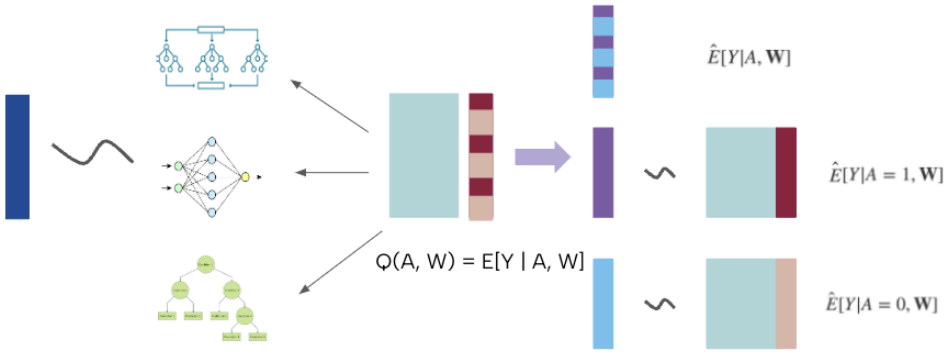


Figure 2: Get Initial Outcome

2.3 Update Initial Outcome

The next step involves modeling the treatment with confounders to estimate the probability under treatment and control. Calculating two inverse probabilities under treatment and control, we generate a clever covariate (CC). This CC is then fitted into a logistic regression, with the only coefficient in

the regression denoted as the "fluaction parameter" (FP). Utilizing the FP and the two inverse probabilities, we update our initial estimate using the formula and the formula are shown as below. This updating step is crucial for making TMLE asymptotically efficient. If either the estimation of the probability of treatment or the mean model from the initial step is accurate, our inference will be correct. This is a really outstanding features for TMLE, called doubly robustness.

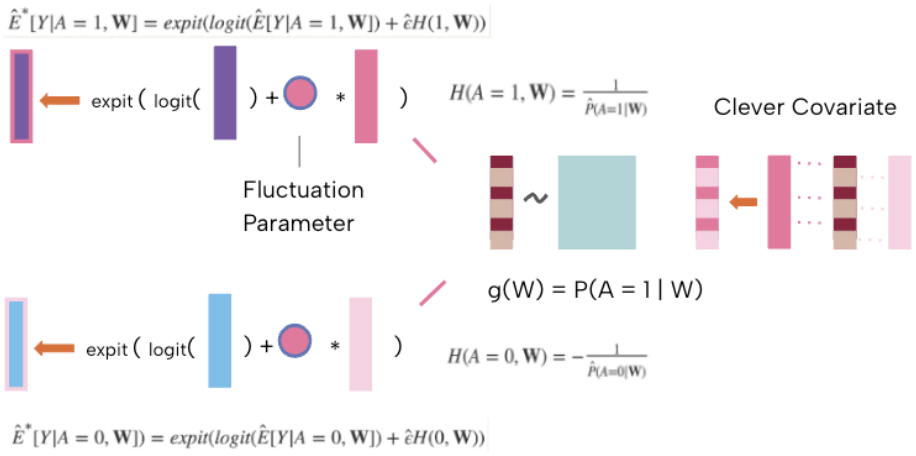
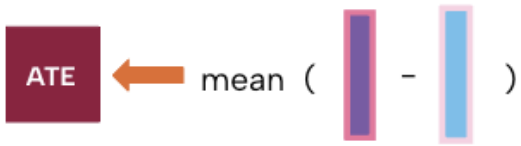


Figure 3: Update Initial Outcome

2.4 Inference

Then with all the updated outcomes settle down, we can do inference on the Average Treatment Effect (ATE), representing the mean effect. The entire TMLE process, with its careful initial estimation, subsequent update, and reliance on double robustness, contributes to a robust inference about the causal relationship between the treatment/exposure variable and the observed outcomes. And its asymptotic normality of Influence Function allows for construction of CI and test.



$$\hat{ATE} = \hat{E}[\hat{E}^*[Y|A=1,\mathbf{W}] - \hat{E}^*[Y|A=0,\mathbf{W}]]$$

Figure 4: Inference on Average Treatment Effect

3 Application

Then with the TMLE algorithm background, we decided to use TMLE to explore the association between music and mental health. The data was collected from Catherine Rasgaitis by a music and mental health survey; and I found that from Kaggle. It contains 33 variables and 735 observations. From the 33 features, we choose 7 variables to do our project. After cleaning the missing values, we left 624 observations. To be clear our the variables in our project, we define our outcome is Music Effect where is binary after transformation (0 - Improved; 1 - Not Improved). The depression level is our exposure which is 0 - 10 levels. Since we want to make it binary, we randomly cut that from 6 since we did not have any source about how the author defined the levels (0 - Not Depression; 1 - Depression). And the confounders include age, hours which people takes to listen to the music, BPM, favorite genre, and anxiety level.

After we did the TMLE in R with random forest, logistic, and lasso machine learning model, our result is shown as below:

Parameter Estimation	0.01546
Estimated Variance	0.0010993
p-value	0.64101
95% CI	(-0.049524, 0.080444)

Table 1: Result From TMLE

From the table, we can see the p-value is actually pretty large. Then we can tell that there is no significant association between the depression level and music effect. One thing to notice from the project is that since we did not deep in to the causal inference this quarter, so the depression is just exposure, but not treatment. And we cannot make the conclusion about causal relationship here.

4 Conclusion

In employing Targeted Maximum Likelihood Estimation (TMLE) to scrutinize the association between depression levels and perceived improvements in music effects, our study harnessed the advantages of TMLE's robustness and adaptability. Despite the non-significant p-value (0.64101) indicating no substantial association, it is crucial to acknowledge the limitations imposed by treating depression as an exposure rather than a treatment, precluding causal conclusions. TMLE, with its strengths in handling complex relationships and machine learning model integration, provided valuable insights into statistical associations. Moving forward, a more in-depth exploration into causal aspects could enrich our understanding of the nuanced dynamics between music and mental health.

5 Reference

1. Hoffman, K. (2020, December 11). An illustrated guide to TMLE, part II: The Algorithm. KHstats. https://www.khstats.com/blog/tmle/tutorial-pt2

2. RASGAITIS, C. (n.d.). Music & Mental Health Survey Results. Kaggle. https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results?select=mxmh_survey_results.csv

3. M Van Der Laan, S Rose Targeted Learning: Causal Inference for Observational and Experimental Data (Springer, New York, 2011).