Estimating SHAP-based Treatment Effects

Kalka, Iris Yacovzada, Nancy

March 2019

1 Introduction

Causal inference from observational data is a crucial an challenging field of work. Often works rely on assessing potential outcomes in order to estimate treatment effects [1]. Many techniques, such as propensity score (PS) techniques, as formalized by Rosenbaum and Rubin [2], model observed data in order to predict treatment effects. These techniques enable controlling confounding in non-experimental studies in medicine and epidemiology.

A central issue researchers are facing using such methods is how to select the set of variables to be included in the model estimating for treatment effect. The bias and variance of the estimated treatment effect can depend strongly on which of these candidate variables are included in the PS model [3].

One assumption in the potential outcomes framework is ignorability, meaning that there are no unmeasured confounders [2]. More specifically, the assumption of ignorability demands that potential outcomes are independent of treatment assignment, conditioned on the set of observed covariates.

A possible limitation of PS methods, as described by Rubin in [4], is that a covariate related to treatment assignment but unrelated to the outcome is handled the same as a covariate related to treatment but also strongly related to outcome. Such inclusion of irrelevant covariates might reduce the efficiency of an estimated treatment effect [5]. In some cases, the use of pretreatment covariates might even increase bias of treatment effect estimators.

Our work is based on an assumption that every covariate can be treated as a covariate effecting both the treatment and the outcome. In this work we wish to examine several approaches for estimating the Average Treatment Effect for Treated (ATT) in studies with observed and unobserved variables that are associated with both treatment and outcome. We believe that during the estimation of the potential outcomes those variables may confound the targeted causal effect. We propose a novel approach to incorporate features importance scores in the construction of the distance matrix in which is used in the matching operation. These scores are based on a method called SHapley Additive exPlanations (SHAP), described in details in [6]. We compere eight different approaches for obtaining pair-wise distances and test them on three types of simulation data.

2 Methods

We compare treatment effect evaluation using several methods. We separate the methods to baseline methods already available in the literature: methods based on an estimator predicting the treatment; methods based on an estimator predicting the outcome; and methods based on two estimators one predicting treatment and the other predicting outcome.

We compare methods by looking at the predicted Average Treatement Effect on Treated (ATT) compared with the true known ATT. Let us recall the definition of ATT:

$$ATT = \mathbb{E}[Y_1 - Y_0|T = 1]$$

For all of these evaluations we used the Scikit-Learn gradient boosting predictors with default parameters [7].

2.1 Baseline

2.1.1 Inverse Propensity Score Weighting

Our first method of estimating ATT is Inverse Propensity Score Weighting (IPW). We use the following equation for calculation of ATT:

$$\begin{aligned} ATT =& \mathbb{E}[Y_1 - Y_0|T = 1] \\ =& \mathbb{E}_X[\mathbb{E}[Y_1 - Y_0|X, T = 1]] \\ =& \mathbb{E}_X[\mathbb{E}[Y_1|X, T = 1] - \mathbb{E}[Y_0|X, T = 1]] \end{aligned}$$

Under strongly ignorable treatment assumption the potential outcomes Y_0, Y_1 are independent of the treatment selection given the observed covariates X. Therefore:

$$ATT = \mathbb{E}_X[\mathbb{E}[Y_1|X, T = 1] - \mathbb{E}[Y_0|X, T = 0]]$$

Thus we use the propensity score and the inverse probability of treatment weighting to compute ATT [8].

2.1.2 Matching

We utilize an additional method which serves as baseline estimation that relies on matching functions. Matching functions exploits known outcomes Y_0 in order to estimate the T_0 of the treated. For every treated individual ($T = 1$) in the cohort, we calculate his potential outcome $\bar{Y_0}$ as follows:

- Find his most similar k untreated individuals ($T = 0$)
- Calculate $\bar{Y_0}$ - the mean outcome of the k closest neighbors

Now, for every treated individual we calculate his treatment effect $\widehat{T}$ by subtracting the mean outcome $\bar{Y_0}$ with the individual's true outcome Y_1. Thus, given a distance matrix between all individuals we use matching to approximate the treatment effect for every treated individual. These estimates allow us

to calculate the total ATT, as explained above, by average of $\widehat{T}$ of all treated individuals'. Distance (and inversely, similarity) between individuals is computed based solely on the propensity score computed for each individual. The distance function we use here is either the Euclidean distance or Mahalanobis distance. We use the Scikit-Learn K-Nearest Neighbors algorithm for identification of the most similar individuals [7].

2.2 Based on treatment prediction

Here we build a single "treatment predictor" $\mathcal{M}_{\mathcal{T}}$: a classifier that uses the $N \times M$ matrix of covariates to predict the treatment. In order to account for the difference in effects of each covariate we chose to look at the features importance obtained based on this model. To make the different features importances comparable we use SHapley Additive exPlanations (SHAP) values [6]. These values are scalar values, explaining the individual contribution of each covariate for each individual in the estimation of the treatment.

Given a matrix of $N \times M$ covariates (where N is the number of individuals and M is the number of covariates) we create a single predictor $\mathcal{M}_{\mathcal{T}}$ from the covariates to the treatment (marked a as a binary value). We then compute the SHAP values of $\mathcal{M}_{\mathcal{T}}$ for every individual resulting in an $N \times M$ matrix S_T.

The distances between individuals are computed based on the SHAP values matrix S, using either Euclidean of Mahalanobis distance for every two individuals. Finally, the ATT is calculated using matching.

2.3 Based on outcome prediction

Here we build a single "outcome predictor" $\mathcal{M}_{\mathcal{Y}}$: a classifier that predicts the outcome given both $N \times M$ matrix of covariates and a boolean vector of length N of the treatment. Based on this prediction model we compute the SHAP values of all input values which yields a $N \times (M + 1)$ matrix.

The distances between individuals are computed based on the SHAP values matrix S, using either Euclidean of Mahalanobis distance for every two individuals. Our computed distance should be reflective only on pretreatment features, regardless if the individual is received treatment or not. Therefore, we remove the SHAP value respective of the treatment assignment, yielding us a matrix S_Y of $N \times M$.

The given matrix is then used to compute distances either through Euclidean distance of Mahalanobis distance between every two individuals. ATT is then calculated using matching based on the distance matrix.

2.4 Based on treatment prediction and outcome prediction

We create two predictors. One predictor $\mathcal{M}_{\mathcal{T}}$ predicts the treatment assignment from covariates. The second predictor $mathcal M_Y$ predicts the outcome from both the covariates and the treatment assignment.

We compute the SHAP values from both treatment and outcome predictions (S_T and S_Y respectively). In order to have comparable SHAP values we remove the column related to treatment from the S_Y matrix.

At this point each individual is represented with two vectors of length M (number of covariates). We then join the vectors by dividing the values in S_Y by those in S_T in an element-wise division. Thus, each individual is represented with a single combined vector of length M.

Distances between individuals are computed using either Euclidean or Mahalanobis distances on combined vectors. ATT is computed using matching based on the distance matrix.

2.5 Simulation

2.5.1 Z-Bias

Recent theoretical studies have shown that conditioning on an instrumental variable (IV), a variable that is associated with the treatment but not associated with outcome except through treatment, can increase both bias and variance of treatment effect estimates. We follow the work from Myers *et al* [9] where they suggested the "Z-Bias simulations", Monte Carlo simulation designed to provide insight into the problem of conditioning on potential IVs and adjusting for instrumental biases. Their results indicated that effect estimates which are conditional on a perfect IV or near-IV may have larger bias and variance than the unconditional estimate. However, they showed that in most scenarios considered, the increases in error due to conditioning were small compared with the total estimation error. Hence, in these cases minimizing unmeasured confounding should be the priority when selecting variables for adjustment, even at the risk of conditioning on IVs. The simulation creates examples where there is: an instrumental variable Z, an unobserved variable affecting both treatment and outcome U, a treatment X and an outcome Y. All four variables are assumed to be binary. The causal graph is described in Figure 1 from the original paper.

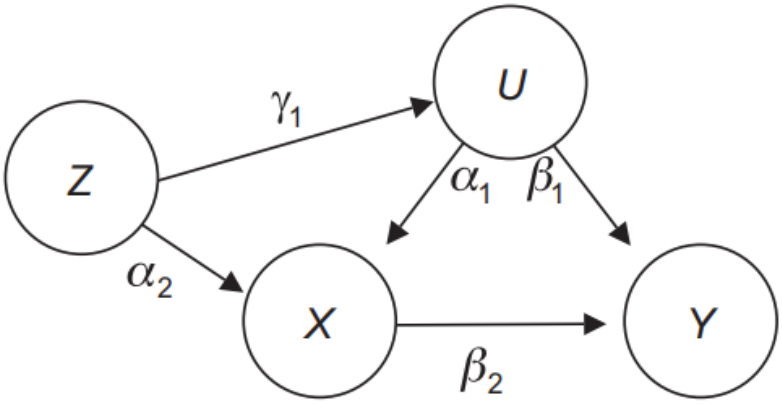

Figure 1: A figure from Myers *et al* depicting the relationships between variables in simulated data.

There are two methods of simulation we use, an additive and a multiplicative version. In both versions the order of variables simulated is consistent in order to ensure that the risk of outcome would depend directly on U and X and indirectly on Z.

In the **additive** version of the simulation the first variable to be simulated is Z that is created with the probability $P(Z = 1) = 0.5$. Next the unobserved variable is simulated so that $P(U = 1|Z) = \gamma_0 + \gamma_1 \cdot Z$. Next the treatment is simulation using $P(X = 1|U, Z) = \alpha_0 + \alpha_1 \cdot U + \alpha_2 \cdot Z$. Last we simulate the outcome using $P(Y = 1|X, U) = \beta_0 + \beta_1 \cdot U + \beta_2 \cdot U$.

In the **multiplicative** version of the simulation the order of simulation is identical to that in the additive version. However, probabilities are computed in a different format as is explained in the following equations:

$$P(Z = 1) = 0.5$$

$$P(U = 1|Z) = \gamma_0 \cdot (\gamma_1^Z)$$
$$P(X = 1|U, Z) = \alpha_0 \cdot (\alpha_1^U) \cdot (\alpha_2^Z)$$
$$P(Y = 1|X, U) = \beta_0 \cdot (\beta_1^U) \cdot (\beta_2^X)$$

The true "exposure effect" and target of estimation is β_2. For simplicity and to reflect a common study framework, all variables are binary. The measured covariate Z may act as a confounder or as an IV for the treatment-outcome pair (X, Y). Note that Z is not a perfect instrument because it is associated with the unobsereved confounder U through γ_1. However, by varying the value of γ_1, we can explore the impacts of conditioning on Z when it is a perfect instrument

and when it is a near-instrument or confounder. A relatively large values of γ_1 can be considered to compare the risks of adjusting for an IV with the benefits of adjusting for a real confounder.

2.5.2 Kang-Schafer

Our first method of simulation is the Kang-Schafer method for creating simulated data [10]. This method assigns four random covariates for every individual (x_i^j $\mathcal{N}(0,1)$ for every $i \in \{1,2,3,4\}$). Propensity is then computed in the following way:

$$p(T=1|X=x) = \frac{1}{1+\exp\left(x_1 - \frac{x_2}{2} + \frac{x_3}{4} + \frac{x_4}{10}\right)}$$

We then assign treatment using the true propensity score as the probability for a binomial variable, yielding treatment assignments that are binary.

Given both the covariates and the treatment assignments, we compute a expected outcome as:

$$\mathbb{E}[Y|X=x,T=t] = 210 + t + 27.4 \cdot x_1 + 13.78 \cdot x_2 + 13.7 \cdot x_3 + 13.7 \cdot x_4$$

Under the strong ignorability assumption and equations from the Kang-Schafer paper we can compute the true ATT using the following steps:

$$\begin{aligned} ATT =& \mathbb{E}[Y_1|X=x,T=1] - \mathbb{E}[Y_0|X=x,T=1] \\ \mathbb{E}[Y_1|T=1,X=x] =& \mathbb{E}[Y|T=1,X=x] \\ \mathbb{E}[Y_0|T=1,X=x] =& \mathbb{E}[Y_0|T=0,X=x] = \mathbb{E}[Y|T=0,X=x] \end{aligned}$$

Therefore $ATT = \mu^{(1)} - \mu^{(0)} = 200 - 220 = -20$

2.6 Results

We simulated data of 10,000 individuals using every time a different simulation approach as explained below. We then use the aforementioned methods to estimate the ATT. For each method we estimated the ATT values using matching on k-neighbors, where $k = 1, 2, ..., 30$.

2.6.1 Z-Bias

First, we simulated data from the causal diagram as presented in Figure 1 with the following coefficients:

- $\gamma_0 = .3$
- $\gamma_1 = .06$
- $\alpha_0 = .3$
- $\alpha_1 = .18$

- $\alpha_2 = .18$
- $\beta_0 = .2$
- $\beta_1 = .36$
- $\beta_2 = .2$

In the case of **Z-bias additive ATT estimation**, we can see that the worst estimation was obtained by the estimations based on two classifiers, the outcome prediction $\mathcal{M}_\mathcal{Y}$ and treatment prediction $\mathcal{M}_\mathcal{T}$. The following estimations resulted identical estimation error, regardless of the distance function, and they also are the best estimations: Matching, SHAP-based treatment prediction $\mathcal{M}_\mathcal{T}$ and SHAP-based outcome prediction $\mathcal{M}_\mathcal{Y}$. The results suggest that there is no improvement in using the SHAP-based distance matrix compered to standard matching with propensity score and in fact the IPW baseline method was the second based estimator. Full results can be found in the IPython notebook.

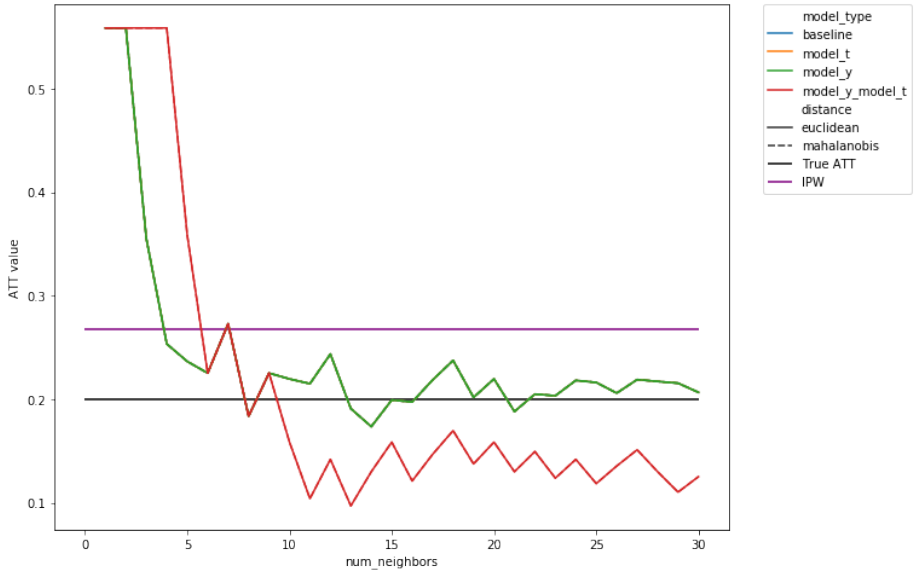

Figure 2: Z bias additive ATT estimations by number of k

In the case of **Z-bias multiplicative ATT estimation**, we can see that the worst estimation was obtained by the IPW baseline method. The following estimations resulted identical estimation error, regardless of the distance function, and they also are the best estimations: Matching, SHAP-based treatment prediction $\mathcal{M}_\mathcal{T}$ and SHAP-based outcome prediction $\mathcal{M}_\mathcal{Y}$. The estimations based on two classifiers, the outcome prediction $\mathcal{M}_\mathcal{Y}$ and treatment prediction $\mathcal{M}_\mathcal{T}$, resulted with higher estimation error than these methods but lower than

IPW. The results suggest that there is no improvement in using the SHAP-based distance matrix compered to standard matching with propensity score. Full results can be found in the IPython notebook.

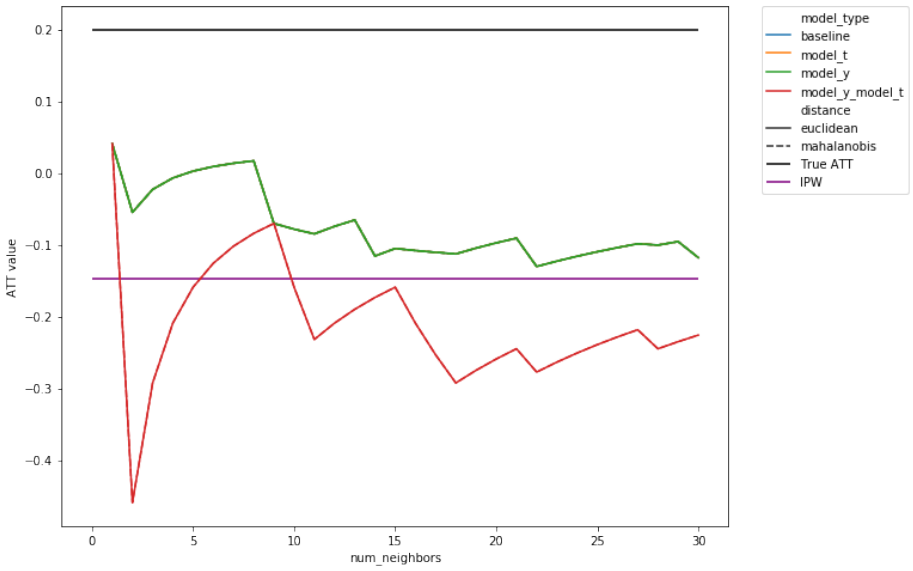

Figure 3: Z bias multiplicative ATT estimations by number of k

Second, we simulated data with the following simulation coefficients:

- $\gamma_0 = .3$
- $\gamma_1 = .06$
- $\alpha_0 = .3$
- $\alpha_1 = .18$
- $\alpha_2 = 0$
- $\beta_0 = .2$
- $\beta_1 = .0$
- $\beta_2 = .2$

Note that here we assigned $\beta_1 = 0$ and $\alpha_2 = 0$ which result that Z is not a IV (we cancelled the arrow from Z to X). In addition, under these coefficients the unobserved variable U doesn't effect the outcome. Results for the additive data are presented in Figure 4.

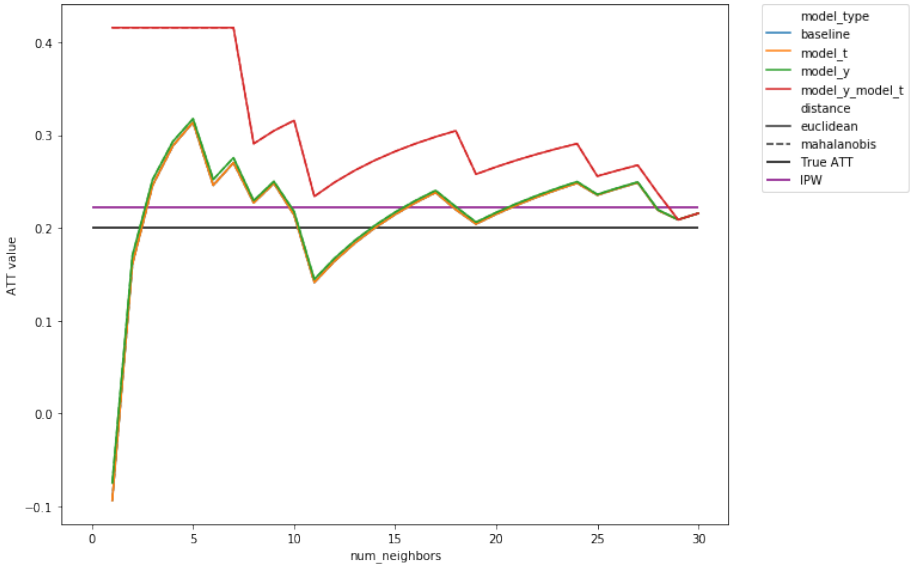

Figure 4: Z bias additive ATT estimations by number of k neighbors - No IV

Results for the multiplicative data are presented in Figure 5.

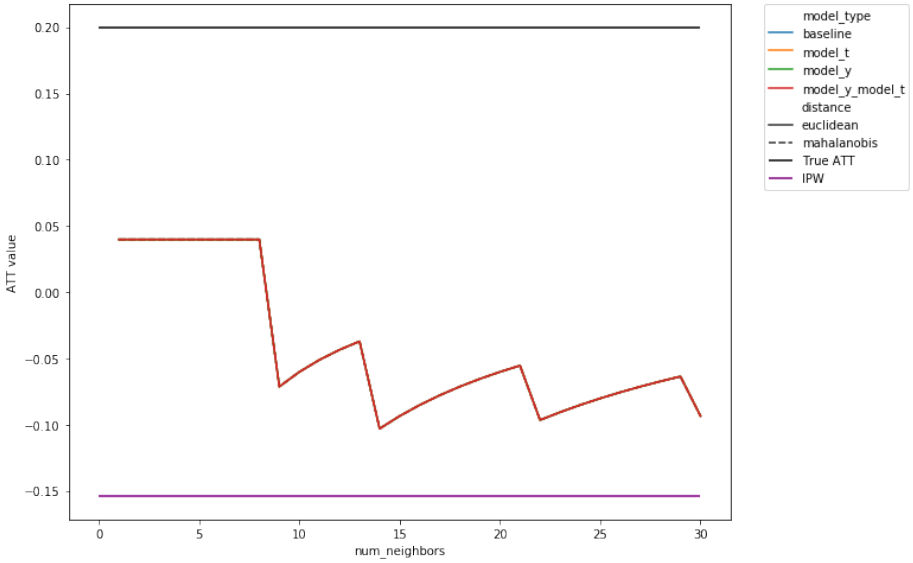

Figure 5: Z bias multiplicative ATT estimations by number of k - No IV

For all models it is shown that both the Euclidean and the Mahalanobis distance functions yielded similar results.

2.6.2 Kang-Schafer method

Results of depicted in Figure 6. We can see that IPW estimation gives an opposite direction of error in estimation compared to all other ATT estimation methods. In edition we can see that ATT estimations based solely on SHAP values from the model predicting treatment yielded the worst results. Following are baseline models and models based solely on the SHAP values from the model prediction the outecome. The best results we given with using SHAP values combination from a model predicting the treatment and a model predicting the outcome. For all models it is shown that both the Euclidean and the Mahalanobis distance functions yielded similar results, thus showing a minimal effect. In addition we can see that although there is some effect on the number of neighbors considered during matching, this effect occurs mostly for very small number of neighbors and is vastly smaller when taking into account more than five neighbors.

We simulated a data of $10,000$ individuals using the Kang-Schafer method explained above. We then use the aforementioned methods to estimate the ATT. For each method we estimated the ATT values using matching on k neighbors, where $k = 1, 2, ..., 30$. Results of depicted in Figure 6.

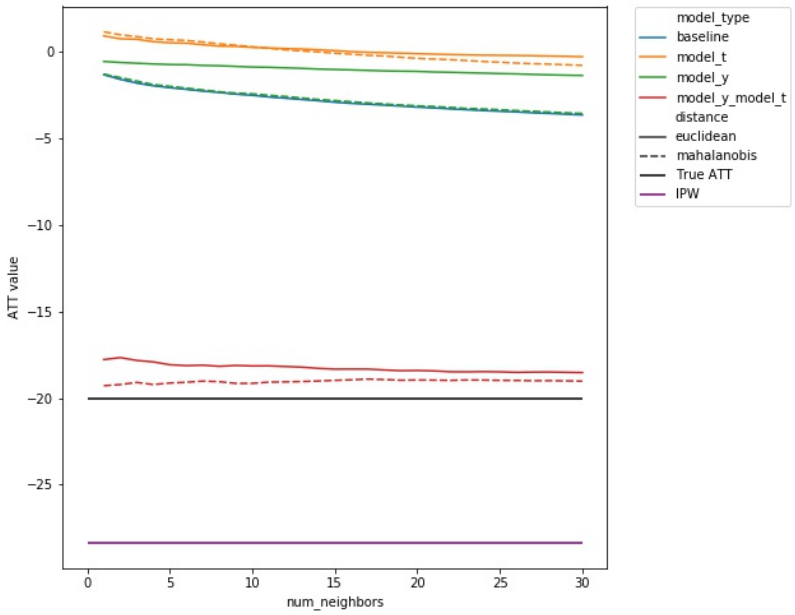

Figure 6: **ATT Estimations for Kang-Schafer Simulation**: This image shows estimated values of ATT on the Kang-Schafer simulated data. True ATT value is -20 and is show in a black solid line. The inverse propensity weighing is marked in a purple solid line. The rest of the estimators are drawn in style according to the distance function used for nearest neighbor matches - Euclidean or Mahalanobis. Color of estimations are saperated by the values used in distance measurements of samples: baseline, based on treatment predicting model, based on outcome predicting model, based on the combination of outcome and treatment prediction models. The x axis is the number of neighbors considered in the nearest neighbor algorithm.

We can see that IPW estimation gives an opposite direction of error in estimation compared to all other ATT estimation methods. In edition we can see that ATT estimations based solely on SHAP values from the model predicting treatment yielded the worst results. Following are baseline models and models based solely on the SHAP values from the model prediction the outecome. The best results we given with using SHAP values combination from a model predicting the treatment and a model predicting the outcome.

For all models it is shown that both the Euclidean and the Mahalanobis distance functions yielded similar results, thus showing a minimal effect.

In addition we can see that although there is some effect on the number of

neighbors considered during matching, this effect occurs mostly for very small number of neighbors and is vastly smaller when taking into account more than five neighbors.

3 Discussion

In the framework of potential outcomes an assumption of ignorability is usually made. This assumption relies on independence between treatment and outcome given covariates. However, reality shows that not all covariates should be considered, as some are in fact instrumental variables effecting only the treatment assignment directly.

Our work aims to create a framework allowing consideration of all covariates while accounting for a difference in effect on treatment and outcome. It proposes a method to improve propensity-based matching by integrating the SHAP algorithm, a method that gives importance score for every variable in the model for every individual in the cohort.

In the case of Z-bias MC simulation our results showed no benefit using the SHAP scores as inout to the distance metric. In addition, both the Euclidean and the Mahalanobis distances yielded the same estimation curves for every k. Surprisingly, the shared consideration of both a model predicting the treatment and a model predicting the outcome didn't overcome the issue of instrumental variables and hence didn't add value compared to using each model independently.

However, for the Kang-Schafer simulation our results show the the shared consideration of both a model predicting the treatment and a model predicting the outcome can indeed overcome issues of instrumental variables. This added value is shown only be relevant when considering both model, where using each model independently has comparable results to baseline estimators.

In addition our work shows little to no impact of the chosen distance function in case of matching when comparing Euclidean and Mahalanobis distances. We remain confident however that matching function based on domain knowledge will give a better estimation of treatment effects.

4 Future Work

The ease of use in our work is meant for cases of high-dimensional data as well as simple data. We believe that the framework created will work well on data with many covariates, however have yet to test it.

Another type of data the framework should be tested on is data with more complex connections. Our current simulations are all based on rather simple relations between variables. A comparison to cases where there are covariate effecting treatment and/or outcome indirectly could give a better understanding of the framework's abilities. This is especially true since the entire framework can also be used on more complex predictors.

Our current framework has only one method of combining the SHAP values of a treatment predicting model and an outcome predicting model. The rational behind using the ratio of SHAP values between the two models is to reduce the impact of covariates that effect the treatment assignment but not the outcome. Although fitting said goal, we hope to try additional methods of combining SHAP values. Specifically we considered using a logarithm of the ration, a ratio of logarithms, an additive factor of the values, etc.

Another field we have yet to explore is the use of the SHAP values in order to perform feature selection on the data. Cases where there are many covariates can at times result in over-fitted models and thus biased predictions of treatment effect. We hope that accounting for the shared effect of a covariate both on the treatment and the outcome could allow to remove uneccessary covariates in a through a causal look.

5 Code

Link to Python code on GitHub

6 References

References

[1] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[2] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[3] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[4] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[5] P. Ding, J. M. Robins, and T.J. Vanderweele. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302, 04 2017.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Younathan Abdia, KB Kulasekera, Somnath Datta, Maxwell Boakye, and Maiying Kong. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5):967–985, 2017.

[9] Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222, 2011.

[10] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.