PROJECT MUSE®

An Application of Matching After Learning To Stretch (MALTS) to the ACIC 2018 Causal Inference Challenge Data

Harsh Parikh, Cynthia Rudin, Alexander Volfovsky

Observational Studies, Volume 5, Issue 2, 2019, pp. 118-130 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2019.0006>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793361/summary>

An Application of Matching After Learning To Stretch (MALTS) to the ACIC 2018 Causal Inference Challenge Data

Harsh Parikh

harsh.parikh@duke.edu

*Department of Computer Science
Duke University
Durham, NC 27710, USA*

Cynthia Rudin¹

cynthia@cs.duke.edu

*Department of Computer Science
Department of Electrical and Computer Engineering
Department of Statistical Sciences
Duke University
Durham, NC 27710, USA*

Alexander Volfovsky¹

alexander.volfovsky@duke.edu

*Department of Statistical Sciences
Duke University
Durham, NC 27710, USA*

Abstract

In the learning-to-match framework for causal inference, a parameterized distance metric is trained on a holdout train set so that the matching yields accurate estimated conditional average treatment effects. This way, the matching can be as accurate as other black box machine learning techniques for causal inference. We use a new learning-to-match algorithm called Matching-After-Learning-To-Stretch (MALTS) (Parikh et al., 2018) to study an observational dataset from the Atlantic Causal Inference Challenge. Other than providing estimates for (conditional) average treatment effects, the MALTS procedure allows practitioners to evaluate matched groups directly, understand where more data might need to be collected and gain an understanding of when estimates can be trusted.¹

Keywords: Matching Algorithm, Causal Inference, Nearest Neighbors

1. Introduction

Matching methods should place “similar” units into matched groups, but the question of whether two units are similar is much more complicated than it might seem. The choice of distance metric used to form the matches can have a large impact on causal conclusions. If we construct a high quality distance metric for matching, not only will we make correct causal conclusions, but we will also have interpretable matches that we can understand and troubleshoot. Examining the matches will allow us to locate possible sources of confounding, areas where treatment and control units do not overlap, or perhaps it will allow us to troubleshoot other problems with our data.

1. Denotes approximately equal contribution

Conversely, a low quality distance metric can lead to poor matches and incorrect conclusions. If the experimenter chooses a fixed distance metric beforehand, such as Euclidean distance (as is typically used in K-nearest neighbors) or logit distance (as in propensity score matching), then it would weigh distances between important covariates equally as compared to the distances between irrelevant covariates. Even with a few irrelevant covariates in the data, this could easily degrade the distance metric far enough that it would affect the treatment effect estimates (Wang et al., 2017). An alternative to Euclidean distance would be to ask experimenters to construct the distance metric using domain expertise, for instance, using a weighted Euclidean distance or by performing coarsened exact matching (Iacus et al., 2011). This, however, requires the user to construct a high-dimensional distance function. Humans are not naturally adept at constructing high dimensional functions manually, and the high degree of freedom in choosing this metric exposes how easily this choice could go wrong.

Ideally, the distance metric should match units together so that *matched groups yield accurate estimated conditional treatment effects*. The *learning-to-match* framework proposed by Parikh et al. (2018) aims to do this. In this framework, the distance metric is trained on a holdout train set. Because the distances are trained, matched groups tend to have more accurate estimated treatment effects. Irrelevant covariates are automatically ignored as part of the learning process, and similarities along relevant covariates tend to be weighted more highly. The impact of arbitrary human choices is reduced, and we can quantitatively judge the quality of the distance metric prior to using it.

Matching After Learning to Stretch (MALTS) (Parikh et al., 2018) is a learning-to-match method that stretches important covariates and shrinks less relevant covariates. By stretching relevant covariates, it forces the distance to be more sensitive to small changes in these covariates. By shrinking less relevant covariates, the distance metric becomes less sensitive to their changes. The stretching and shrinking parameters in the distance metric are learned from a holdout train set, as a special case of learning the parameters of a diagonal Mahalanobis distance matrix. If the Mahalanobis distance is not forced to be diagonal, it would induce more general distances, including rotations, but in this work, we consider only stretching and shrinking of individual covariates for interpretability. MALTS handles categorical covariates by exact matching on as many relevant categorical covariates as possible, using ideas from the FLAME algorithm of Wang et al. (2017).

We applied the new learning-to-match framework on the Empirical Investigation of Methods for Heterogeneity Workshop data from the 2018 Atlantic Causal Inference Conference (ACIC). The ACIC dataset emulates the data and intervention in the National Study of Learning Mindsets. It include school identifiers, self-reported expectation for success in future (S_3), race (C_1), gender (C_2), first-generation status (C_3), level of urbanicity (X_C), school-level mean mindset (X_1), school achievement level (X_2), racial/ethnic composition of school (X_3), poverty concentration of school (X_4), and school size (X_5). Outcome is a continuous measure of student achievement (Y), and treatment is a mindset intervention indicated in the data by T . We find that even though MALTS induces only simple stretching and shrinking of the covariates, it tends to produce results that are similar to a black box machine learning methods for treatment effect estimation like BART (Chipman et al., 2010), Causal Forest (Wager and Athey, 2017) or CFR (Johansson et al., 2016). Moreover, it produces matched groups that we can display, critique, and troubleshoot before estima-

tion takes place. The matched groups are not created posthoc to explain a black box, they are created as part of an interpretable, auditable process.

Briefly, our findings using *Matching After Learning to Stretch* (MALTS) methodology highlights the importance of self-reported expectations for success in the future in determination of the outcome for both treatment and control sets jointly. We also observe the hints of goldilock’s effect for covariates regarding school-level achievements. The average trend observed jointly on the self-reported expectation for success in the future and urbanicity shows that higher values of both are possibly correlated with higher treatment effects. We find interesting behavior for level 3 of urbanicity that is different from other levels of urbanicity. Finally, we only observed small heterogeneity in treatment effects across gender and almost no heterogeneity for first-generation status. We present these results as an analysis pipeline—we first assess the quality of matched groups and then provide estimates that can account for that quality.

2. Methodology and Motivation

Matching After Learning to Stretch (MALTS) is a matching method that estimates conditional average treatment effects (CATEs) (defined as difference of outcome under treatment, $Y^{(t)}$ and outcome under control, $Y^{(c)}$ for a given x , mathematically formulated as $E[(Y^{(t)} - Y^{(c)})|X = x]$) using K-nearest neighbors (KNN) for estimating the counterfactual outcomes conditional at a given location in covariate space, with a learned distance metric (Parikh et al., 2018). Because the distance metric is learned from a training set, MALTS, and its discrete counterpart “Fast Large-scale Almost Matching Exactly Approach to Causal Inference” (FLAME), are part of the learning-to-match framework (Wang et al., 2017; Dieng et al., 2019). MALTS assumes that there is no unobserved confounding.

MALTS learns a distance metric L such that the following minimization problem is approximately solved on the normalized training set. The train set is normalized such that each covariate has zero sample mean and unit sample variance. The right side of the optimization equation accounts for the aggregate error in estimation of y_i ’s as an average of outcome y_k of its k-nearest neighbors according to the learned distance metric.

$$L \in \operatorname{argmin}_{L'} \left(\sum_{S \in \{C, T\}} \sum_{i \in S} \left(y_i - \frac{1}{K} \sum_{k \in KNN(y_i, L') \subseteq S} y_k \right)^2 \right) \quad (1)$$

The learned distance metric forces the samples with similar outcomes in the set to be closer in covariate space. In this implementation, MALTS’ distance metric for KNN is parameterized by a matrix, termed as L , that handles continuous and discrete variables differently. Let the subscript c indicate continuous variables and d indicate discrete variables then the distance metric is:

$$\begin{aligned} \text{distance}_L(a, b) &= d_{L_c}(a_c, b_c) + d_{L_d}(a_d, b_d), \text{ where } L = [L_c, L_d] \\ \text{and } d_{L_c}(a_c, b_c) &= \|L_c a_c - L_c b_c\|_2^2, \quad d_{L_d}(a_d, b_d) = \sum_{j=0}^{n_d} (L_d^{j,j})^2 1[a_{d(j)} \neq b_{d(j)}]. \end{aligned}$$

MALTS learns the distance metric parameter L using a non-gradient based optimization method with the help of python3’s scipy package (Jones et al., 2001).

The learned distance metric is used for matching the units in the estimation set. In principle the proposed procedure allows for the full data to be used for estimation of CATEs. However, it might be preferable to keep the two sets completely separated. This separation allows for the training to inform additional data collection (for example, if in the training set we find a few nearest neighbors who form very tight groups while the others form loose groups then we might want to explore the space of the loose groups to see if additional data might lead to tighter matches). For each unit, we find both the k nearest neighbors in the control group (KNN_c) and the k nearest neighbors in the treatment group (KNN_t). We estimate the outcome under control $\hat{y}_i^{(c)} = \frac{1}{K} \sum_{k \in KNN_c} y_k^{(c)}$ and the outcome under treatment $\hat{y}_i^{(t)} = \frac{1}{K} \sum_{k \in KNN_t} y_k^{(t)}$. (The superscript (t) refers to the unit in treated set and the superscript (c) refers to the unit in control set.) Thus we estimate the conditional average treatment effect (CATE) as the expected difference between outcomes for units in treated set and control set conditional on given covariates using the relation:

$$E[Y^{(t)} - Y^{(c)} | X = x_i] = \hat{y}_i^{(t)} - \hat{y}_i^{(c)}.$$

We are able to use MALTS to identify outliers or low-quality matches by examining the diameter of the matched group. The diameter provides an intuitive comparison between matched groups: Matched groups that span a large portion of covariate space tend to be low-quality matches. If desired, one can prune the matched groups of large diameter. The diameter of the matched group $MG(x_i)$ for query unit x_i is defined as $\text{diameter}(x_i) = \max_{j \in MG(x_i)} \text{distance}_L(x_i, x_j)$ where matched group $MG(x_i)$ is formed by a union of $KNN_c(x_i)$ and $KNN_t(x_i)$. Instead of pruning, one can create an aggregate estimator by weighting each group inversely proportional to a function of its diameter.

3. Workshop Results

We presented results for a preliminary version of the matching algorithm during the ACIC challenge workshop. In this preliminary approach, we embedded the discrete covariates in Euclidean space and learned a single distance metric for the combination of continuous and embedded (discrete) covariates. This approach yielded interpretable matched groups, and we learned from this exercise that it is important to match on a student’s self reported expectations of future success (S_3) and probably less important to match on gender and first generation status. However, the embedding of the discrete variables is somewhat misleading within our distance metric learning framework. Specifically, it can make units appear artificially similar if they both belong to rare groups. To address this challenge in the post-workshop analysis, we considered a distance metric that explicitly penalizes non-exact matching on discrete covariates without requiring an apriori embedding of the covariates in Euclidean space, which we described above. We note that generalizations of this method (discussed in Parikh et al., 2018) allow for flexible data driven embeddings of complex data that complement the distance metric learning problem. Distance metric learning for nearest neighbors has also been considered in non-causal settings (Goldberger et al., 2005).

4. Post Workshop Analysis

We estimate treatment effects by using the Matching-After-Learning-to-Stretch (MALTS) methodology to construct matched groups and evaluate conditional average treatment effects. Covariates in the ACIC dataset include a school identifier, self-reported expectation for success in future (S_3), race (C_1), gender (C_2), first-generation status (C_3), level of urbanicity (X_C), school-level mean mindset (X_1), school achievement level (X_2), racial/ethnic composition of school (X_3), poverty concentration of school (X_4), school size (X_5), outcome variable (Y), and treatment indicator (T). For each unit we have continuous measurements $X_1, \dots, X_5$ and discrete covariates S_3, C_1, C_2, C_3 and X_C , a continuous outcome variable Y and a binary treatment indicator T .

4.1 Learning a Distance Metric and Match Group Quality Analysis

We use a distance metric of the form defined in Equation (2) below, where MALTS learns a scaling on the exact matching distance for discrete covariates and stretch on the Euclidean distance for the continuous covariates:

$$\begin{aligned} \text{distance}_L(a, b) = & L_{1,1}^2 \mathbf{1}[S_3(a) \neq S_3(b)] + L_{2,2}^2 \mathbf{1}[C_1(a) \neq C_1(b)] + L_{3,3}^2 \mathbf{1}[C_2(a) \neq C_2(b)] + \\ & L_{4,4}^2 \mathbf{1}[C_3(a) \neq C_3(b)] + L_{5,5}^2 \mathbf{1}[X_C(a) \neq X_C(b)] + \\ & \sum_{j=1}^5 L_{j+5,j+5}^2 (X_j(a) - X_j(b))^2. \end{aligned}$$

MALTS requires a training set and an estimation set and so we randomly partition the data into these two components: 5% is reserved for training, while the rest is used for estimation of conditional average treatment effects (CATEs). MALTS uses a non-gradient based optimization method to find the optimal L that minimizes (or approximately minimizes) the quadratic loss in Equation (1). The optimum for one run is presented in Table 1 with each element of the vector describing the relevant diagonal entry. Exact matching on S_3 appears to be important, as we can interpret from the learned L and the distance metric, a mismatch as incurring a cost of at least 17.81 (equal to $L_{1,1}^2$)—in comparison, being matched to an individual of the wrong race or ethnicity only costs 1.21 (equal to $L_{2,2}^2$). These costs translate directly to the quality of matches that we can find for each unit in our estimation set. This means S_3 is important in determining the outcomes for control unit Y_c and the outcome for treated unit Y_t jointly.

Table 1: *Stretch Values*: Diagonal entries of the L matrix learned using MALTS.

	S3	C1	C2	C3	XC	X1	X2	X3	X4	X5
Index in L matrix	(1,1)	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(7,7)	(8,8)	(9,9)	(10,10)
Corresponding L value	4.22	1.1	0.65	0.04	2.78	2.56	0.32	1.92	0.64	1.83
Corresponding L^2 value	17.808	1.210	0.422	0.002	7.728	6.554	0.102	3.686	0.409	3.349
Relative importance (value/max_value)*100	100	26.02	15.39	0.89	65.87	60.52	7.52	45.52	15.22	43.24

Figure 1 (a) shows that as the level of S_3 increases, so does the value of Y . To evaluate whether we have sufficient data to make statements about conditional treatment effects for the different levels of S_3 we consider the distribution of diameters of matched groups

Figure 1: *Trends for S_3 and Diameter*: (a) Variation in values of Y for different levels of self-reported expectation of success (S_3). (b) Variation in diameter of matched groups for different levels of self-reported expectation of success (S_3). Levels 1 and 2 do not have high-quality matched groups. (c) Histogram of the diameter of matched groups in estimation set. Groups whose diameter is too large could potentially be omitted from the analysis.

for each of the levels of S_3 . Again, diameter is defined as the maximal distance to one of the nearest neighbors for each unit. Figure 1 (b) demonstrates that individuals at Levels 1 and 2 of S_3 are generally not well matched. While it is true that these are the lowest frequency levels of S_3 , the matched diameter provides a deeper insight: We are likely to have bad matches if we fail to match exactly on them, while if we match exactly on these levels we appear to be paying a large penalty for failing to match on other covariates. This is evidence of non-overlap and in our analysis, it is important to prune or down-weight such bad matched groups to avoid poor performance of the estimator. Figure 1(c) shows the histogram of diameters of all matched groups for units in our estimation set. If we take the tightest 75% or 50% of the matched groups then we need to prune groups with diameter greater than 2.79 or 1.21 respectively.

4.2 Heterogeneity of Treatment Effect

For different levels of pruning based on the diameter of matched groups, Figure 2 shows the variation of CATEs for different levels of S_3 . Analyzing Figure 2 (a), it appears there is an initial decrease and then increase in average CATEs for different levels of S_3 , however from the match quality analysis we know that matches for Level 1 and Level 2 are not very reliable, so this non-monotonicity could be misleading. Figure 2 (b) shows the trend after we prune at the 75th percentile, where Levels 1 and 2 have been completely removed from the matched group analysis. We see that the trend in the CATEs across levels of S_3 dissipates. A similar analysis can also be performed for the 50th percentile pruned dataset as shown in Figure 2 (c).

Following a similar evaluation as that for S_3 we observe that exactly matching on urbanicity (X_C) appears important. Since S_3 and X_C have many levels, we present the joint variability of CATEs in Figures 3 (a), (b) and (c) after no pruning on the diameter, pruning at the 75% level and pruning at the 50% level. It is evident that pruning removes spurious matched group allowing us to evaluate underlying trends. One trend we can infer from the plot is that a low level of urbanicity (X_C) and a high level of self-reported expectation

Figure 2: *CATEs for different values of S_3* : Variation in predicted CATEs for different levels of self-reported expectation of success (S_3) for (a) all the samples in estimation set, (b) for samples beneath the 75th percentile of the estimation set based on the diameter of the matched group, (c) for samples beneath the 50th percentile of the estimation set based on the diameter of the matched group.

Figure 3: *Treatment Effects with respect to S_3 and X_C* : Variation in predicted CATEs for different levels of self-reported expectation of success (S_3) and urbanicity (X_C) for (a) all the samples in the estimation set, (b) for samples beneath the 75th percentile of the estimation set, (c) for samples beneath the 50th percentile of the estimation set based on the diameter of the matched group.

(S_3) tends to have on an average higher individual treatment effects. We also note that marginally, urbanicity Level 3 exhibits substantially lower CATEs than the rest of the levels.

4.2.1 WHO DID WE MATCH?

Table 2 (a), (b), and (c), highlights three specific examples of a good, a bad and an “ugly” matched group. The good matched group is tight, with diameter equal to zero. The bad matched group is interesting because for most of the units it is trying to force a match on S_3 , and hence not making good matches on other covariates. The ugly one is unable to find any sample similar to the query sample. During estimation the latter two matched groups may be pruned as the poor quality matches are likely to produce low quality estimates.

Table 2: *Example Matched Groups- The Good, The Bad, The Ugly*: (a) Example of a good matched group produced by MALTS with diameter equal to zero (exact match on all covariates). For the query point, MALTS finds 10 nearest neighbors in the control group and 10 nearest neighbors in the treatment group, all of whom are exactly matched. We report only the averaged Y_c and Y_t over all 10 matches for each group respectively. (b) Examples of a bad and (c) an ugly matched group produced by MALTS.

The Good, diameter: 0.0													
S3	C1	C2	C3	XC	X1	X2	X3	X4	X5	$\hat{Y}$	T		
6	4	2	0	4	-1.1	0.7	-0.1	1.3	1.9				
6	4	2	0	4	-1.1	0.7	-0.1	1.3	1.9	-0.01	0		
6	4	2	0	4	-1.1	0.7	-0.1	1.3	1.9	0.2	1		

The Bad, diameter: 19.1													
S3	C1	C2	C3	XC	X1	X2	X3	X4	X5	$\hat{Y}$	T		
1	1	1	1	1	1	1	1	1	1	1	1	1	4
14	4	14	1	2	4	14	4	12	13	14	4	13	4
1	2	1	1	2	1	1	1	2	1	1	1	2	1
1	0	0	1	0	1	1	0	0	1	0	1	1	1
2	2	0	2	3	2	2	4	3	3	1	2	3	2
0.4	0.4	0.2	0.3	0.1	-0.2	0.2	0.2	0.4	0.4	0.1	0.4	0.2	1
-0.2	-0.2	-0.8	0.6	-0.3	-1	-1.1	-0.7	-1.1	-0.9	-1.1	-0.9	-1.3	-0.2
-1.1	-1.1	-1	-1.3	-1	-1.1	-1	-1.1	-1.1	-1.1	-1.1	-1.1	-1.1	-1.1
-1.2	-1.2	-0.1	0.2	-0.8	0	-1.2	-1.2	-0.8	0.1	0	-1.4	-1.2	-0.1
-0.6	-0.6	-0.6	-0.4	0.2	0.3	-0.8	0.3	-0.4	-0.6	-0.3	-1.1	-0.4	-0.6
	-0.8	-0.8	-0.5	-1.6	-1.1	0.3	-0.7	-0.8	0	-0.4	0.7	-1.6	-0.8
	0	0	0	0	0	0	0	1	1	1	1	1	1

The Ugly, diameter: 21.5													
S3	C1	C2	C3	XC	X1	X2	X3	X4	X5	$\hat{Y}$	T		
2	2	2	2	2	5	2	2	6	5	4	5	5	5
13	4	14	12	13	4	4	13	4	15	4	4	1	4
1	1	2	1	1	1	1	2	1	1	2	1	2	1
1	1	1	1	1	1	1	0	1	1	1	1	1	0
1	1	3	1	1	1	1	1	1	1	1	1	1	1
1.7	1.7	0.9	1.7	1.7	1.7	1.7	0.9	1.7	1.7	1.7	1.7	1.7	1.7
-0.2	-0.2	0.1	-0.2	-0.2	-0.2	-0.2	-1	-0.6	-0.6	-0.2	-0.2	-0.2	-0.6
-1.5	-1.5	-0.6	-1.5	-1.5	-1.5	-1.5	-1.2	-1	-1.5	-1	-1.5	-1	-1
-0.9	-0.9	0.1	-0.9	-0.9	-0.9	-0.9	0.7	0.5	0.5	-0.9	-0.9	0.5	0.5
-1.7	-1.7	-1.1	-1.7	-1.7	-1.7	-1.7	-1.2	-1.6	-1.6	-1.7	-1.7	-1.6	-1.6
	-0.3	0	-0.8	-0.7	-0.3	-0.6	-0.9	-0.5	0.4	-0.2	-0.6	-0.3	-0.8
	0	0	0	0	0	0	0	1	1	1	1	1	1

4.2.2 EVALUATING THE TREATMENT EFFECTS

The average treatment effect (ATE) (difference of outcome under treatment, $Y^{(t)}$ and outcome under control, $Y^{(c)}$, that is, $E[(Y^{(t)} - Y^{(c)})]$ estimated above is not sensitive to pruning or weighting of the matched groups based on the diameter of the matched group. Table 3 illustrates the same behavior over three different cases: two different levels of pruning and an exponential weighting of the matched groups. The ATE is always estimated in the interval of 0.257 to 0.261.

Table 3: Average Treatment Effect (ATE) estimates on the full estimation set, 75th percentile estimation set, 50th percentile estimation set and by exponential weighting each matched group in the estimation set with $w_i = e^{-0.1 \times \text{diameter}(x_i)}$.

	Estimated ATE ($\hat{\mu}$)	Standard Error ($\frac{\hat{\sigma}}{\sqrt{n}}$)
<i>Unpruned</i>	0.2570	0.00253
<i>75th Percentile</i>	0.2604	0.00292
<i>50th Percentile</i>	0.2599	0.00353
<i>Exponentially Weighted</i>	0.2575	0.00219

Following our analysis of matched groups, we investigate the “Goldilock’s” hypothesis on the X_2 variable: that the treatment effect is higher for an average school level achievement compared to low or high achievement. Figure 4 (a) shows that the match quality for low levels of school achievement is poor and in fact, it might be hard to make meaningful statements about matched groups with diameter more than 4. While eliminating these bad matched groups allows us to comment on X_2 , we note from Figure 4 (b) that this pruning removes all large and small values of X_1 (mean fixed mindset). This means that we can primarily comment about a school with middle-mindset individuals with medium to high achievement levels. This is a slight restatement of the Goldilocks’s hypothesis as it contains conditions on the medium mindset. In Figure 5 (a) we can observe that there is a peak near $X_2 = 1$ approximately and it decreases as we move away in either direction, i.e., towards $X_2 = 2$ or $X_2 = 0$. This behavior is consistent for both the unpruned and pruned versions. The pruned version however is much smoother and shows that a Goldilock’s effect is potentially present, but it may not be large.

We also analyze the variability in CATEs as a function of X_1 : we observe a sharp transition in trends for schools with middle-level mindset. We notice for middle-level mindset schools, the treatment effect decreases as the mindset increases. Evaluating X_1 and X_2 jointly, Figure 5 (c) shows that for high values of both X_1 and X_2 , the variability in treatment effect appears to be large, but this can be due to lack of data at the extremes; when instead considering middle-values for both covariates, treatment effects are more stable.

Lastly, we evaluate the marginal treatment effect for different levels of categorical variables for race (C_1) and urbanicity (X_C). As we can infer from Figure 6 (a), the average diameters for all levels are almost the same, so pruning diameters greater than 4 will still yield enough samples to analyze. The average CATEs for all other levels are similar to the ATE or slightly higher, while the average CATE for individuals at Level 3 is lower. This can potentially indicate an effect involving urbanicity at Level 3. Figure 7 shows the marginal treatment effect for gender (C_2) and first-generation status (C_3). There is little variability

Figure 4: *Diameter variation with respect to (a) X_2 and (b) X_1 .* Variation in diameter of matched groups for (a) different values of school level achievement (X_2), (b) different values of mean fixed-mindset (X_1). The black curves shows the trend curves fitted with support vector regression.

Figure 5: X_1, X_2 versus *CATEs*: Variation in predicted CATEs for different values of (a) school level achievement (X_2), for all the samples in the estimation set and for the samples after pruning for matched groups with diameter exceeding 4. Both the blue and the orange curves are fitted using support vector regression with gaussian kernel to the CATE as a function of X_2 ; (b) school level students' fixed mindset (X_1), for unpruned estimation set. The black curve is fitted using support vector regression with gaussian kernel to the CATE as a function of X_1 . (c) Contour plot of variation in predicted CATEs for different values of school level students' fixed mindset (X_1) and school level achievements (X_2) for all matched groups.

Figure 6: *Trends for X_C* : Variation in (a) diameter of matched groups, (b) predicted CATEs, for different levels of urbanicity (XC) after pruning groups with diameter more than 4.

Figure 7: *Trends for C_2 and C_3* : Variation in CATEs for (a) different levels of gender in the full estimation set, (b) different levels of gender in the estimation set pruned at diameters greater than 4, (c) different levels of first generation status in the full estimation set, (d) different levels of first generation status in the estimation set pruned at diameters greater than 4.

in average CATEs across the first generation statuses while the average CATEs for gender Level 2 is on an average higher than average CATEs for gender Level 1.

5. Conclusion and Discussion

In a nutshell, the average treatment effect estimated by MALTS is approximately 0.25 and there are some covariates that moderate this effect. For urbanicity, X_C , we see a different behavior specifically for Level 3 while for other levels, the estimates are fairly consistent. Since matching on S_3 appears important, we probably should collect additional data for lower levels of self-reported expectations. Alternatively, it is possible that the scale for this variable simply needs to be recalibrated; perhaps Levels 1 and 2 are not meaningful.

The school level achievement covariate (X_2) shows signs of Goldilocks’s effect and should be further tested, while the covariate concerning the school-level mean of students’ fixed mindset (X_1) shows non-monotonicity in the middle region of its distribution. Finally, the treatment effect seems to vary slightly across genders and has almost no variation with respect to first-generation status.

MALTS permits several important aspects of our analysis to be made easier: understanding importance of different covariates, the underlying trends in heterogeneity of treatment effects, and interpretation of the trends by analyzing matched groups. It also provides confidence on the predictions of treatment effects based on the diameter of the matched groups. In certain situations like one shown in Figure 5, pruning of poor quality matches yields more stable results. We also observe that ATE prediction by MALTS for this dataset seems to be fairly robust to analyst choices.

Acknowledgments

This work supported by DHHS, PHS, NIH, and NIBI under grant 1R01EB025021-01, and also by the Duke Energy Initiative.

References

- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, pages 266–298.
- Dieng, A., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2019). Almost-exact matching with replacement for causal inference. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*.
- Goldberger, J., Hinton, G. E., Roweis, S. T., and Salakhutdinov, R. R. (2005). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, pages 513–520.
- Iacus, S. M., King, G., and Porro, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, page mpr013.
- Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2018-11-10]. Available from: <http://www.scipy.org/>.

- Parikh, H., Rudin, C., and Volfovsky, A. (2018). Malts: Matching after learning to stretch. *ArXiv e-prints arXiv:1811.07415*.
- Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.
- Wang, T., Roy, S., Rudin, C., and Volfovsky, A. (2017). FLAME: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*.