MALTS: Matching After Learning To Stretch

Harsh Parikh, Cynthia Rudin, Alexander Volfovsky

October 18, 2019

Abstract

We introduce a flexible framework that produces high-quality almost-exact matches for causal inference. Most prior work in matching uses ad-hoc distance metrics, often leading to poor quality matches, particularly when there are irrelevant covariates. In this work, we learn an interpretable distance metric for matching, which leads to substantially higher quality matches. The learned distance metric stretches the covariates according to their contribution to outcome prediction. The framework is flexible in that the user can choose the form of the distance metric and the type of optimization algorithm. Our ability to learn flexible distance metrics leads to matches that are interpretable and useful for the estimation of conditional average treatment effects.

Keywords: causal inference, matching, nearest neighbor, distance metric learning

1 Introduction

Matching methods are used throughout the social and health sciences to make causal conclusions where access to randomized trials is scarce but observational data are widely available. Matching methods construct groups of similar individuals, some of whom select into treatment and some of whom select into control, allowing for direct comparison of outcomes between these populations. Matching methods are particularly interpretable since they allow fine-grained troubleshooting of the data. For instance, examining the matched groups may allow the user to detect unmeasured confounding that led some units to have a higher chance of being treated or a higher chance of leading to a positive outcome. Having high-quality matches also allows the user to estimate nonlinear treatment effects with lower bias than parametric approaches. The quality of the matches is our main consideration in this work.

Typically, matching methods place units that are close together into the same matched group, where closeness is measured in terms of a pre-defined distance (e.g., exact, coarsened exact, Euclidean, etc.), while maintaining balance constraints between treatment and control units. Despite its merits, this classical paradigm has flaws, namely that it relies heavily on a prespecified distance metric. The distance metric cannot be determined without an

understanding of the importance of the variables; for instance, the quality of matches for any prespecified distance weighing all covariates equally will degrade as the number of irrelevant covariates increases. This is true irrespective of the matching methodology employed. This has previously been referred to as the toenail problem (Wang et al. 2017, Dieng et al. 2019), where the inclusion of irrelevant covariates (like toenail length) with nonzero weights can overwhelm the metric for matching. A related concern is that the covariates may be scaled differently, where a given distance along one covariate has a different impact than the same distance along a different covariate; in this case, if the weights on the covariates are chosen poorly, the total distance metric can inadvertently be determined by less relevant covariates, again leading to lower quality matches.

Ideally, the distance metric would capture important covariates that significantly contribute in generating the outcome, so that after matching, treatment effect estimates computed within the matched groups would be accurate estimates of treatment effects. If the researcher knows how to choose the distance metric so that it yields accurate treatment effect estimates, this would solve the problem. However, there is no reason to believe that this is achievable in high-dimensional and complex data settings. Producing high dimensional functions to characterize data is a task at which humans are not naturally adept.

In this work, we propose a framework for matching where an interpretable distance measure between matched units is learned from a held-out training set. As long as the distance metric generalizes from the training set to the full sample, we are able to compute high-quality matches and accurate estimates of conditional average treatment effects (CATEs) within the matched groups. One can use any form of distance metric to train, and in this work, we focus on exact matching for discrete variables and generalized Mahalanobis distances for continuous variables. By definition, the generalized Mahalanobis distance is determined by a matrix. If the matrix is diagonal, the distance calculation represents a stretch for each covariate. Irrelevant covariates will be compressed so that their values are always effectively zero. Highly relevant covariates will be stretched so that for two units to be considered a match, they must have very similar values for those covariates. In this way, diagonal matrices lead to very interpretable distance metrics. If the Mahalanobis distance matrix is not constrained to be diagonal, then it induces a stretch and rotation, leading to more flexible but less interpretable notions of distance.

The new framework is called Learning-to-Match, and the algorithm introduced in this work is called Matching After Learning to Stretch (MALTS). We tested MALTS against several other matching methods in simulation studies, where ground truth CATEs are known. In these experiments, MALTS achieves substantially and consistently better results than other matching methods including Genmatch, propensity score matching, and standard (non-learned) Mahalanobis distance in estimating CATEs. Even though our method is heavily constrained to produce interpretable matches, it performs at the same level as non-matching methods that are designed to fit extremely flexible but uninterpretable models directly to the response surface.

2 Related work

Since the 1970's, the causal inference literature on matching methods has been concentrated on dimension reduction techniques (e.g., Rubin 1973a,b, 1976, Cochran and Rubin 1973). In this literature, the leading approach for dimension reduction uses the propensity score, the conditional probability of treatment given covariate information. Propensity score methods target average treatment effects and so do not produce exact matches or almost-exact matches. When treatment is binary, they project data onto one dimension, and closeness of units in propensity score does not imply their closeness in covariate space. As a result, the matches cannot directly be used for estimating heterogeneous treatment effects. Regression methods can be used for CATE estimation, but this assumes that the regression method is correctly specified – or in the case of doubly robust estimation (e.g., Farrell 2015) either the propensity model or the outcome model needs to be correctly specified. Machine learning approaches generalize regression approaches and can create models that are extremely flexible and predict outcomes accurately for both treatment and control groups (Hill 2011, Chernozhukov et al. 2016, Hahn et al. 2017). However, complicated regression methods lose the interpretability inherent to almost-exact matches. Analogously, there is substantial work on learning distance metrics (e.g., Goldberger et al. 2005) again leading to a sacrifice in interpretability. In practice, MALTS performs similarly to (or better than) several machine learning methods in our experiments, despite being restricted to interpretable almost-exact matches with an interpretable distance metric.

A flexible setup for producing high-quality matches is provided by the optimal matching literature (Rosenbaum 2016). These are built on network flow algorithms and integer programming to produce matches that are constrained in user-defined ways (Zubizarreta 2012, Zubizarreta et al. 2014, Keele and Zubizarreta 2014, Resa and Zubizarreta 2016, Noor-E-Alam and Rudin 2015b,a, Kallus 2017). In all of these approaches, the user defines the distance metric rather than learning it through data, which is time-consuming and likely inaccurate, potentially leading to poor quality of the matched groups.

An alternative to optimal matching is coarsened exact matching (CEM) (Iacus et al. 2012), an approach that requires users to specify explicit bins for all covariates on which to construct matches. This requires users to know in advance that the outcomes are insensitive to movements within many high-dimensional bins, which is essentially equivalent to the user knowing the answer to the problem we investigate in this work. Large amounts of user choice to define these bins can also lead to unintentional user bias. By *learning* the stretching rather than asking the user to define it as in CEM, this bias is potentially reduced. The present work builds on work of Wang et al. (2017), Dieng et al. (2019) where a discrete distance metric is learned by considering the prediction quality of the covariate sets.

MALTS was used for the Atlantic Causal Inference Competition (Parikh et al. 2019).

3 Learning-to-Match Framework

Within this framework, we perform treatment effect estimation using following three stages: 1) learning a distance metric, 2) matching samples, and 3) estimating CATEs.

We denote the p dimensional covariate vector space as $\mathcal{X} \subset \mathbb{R}^p$ and the unidimensional outcome space by $\mathcal{Y} \subset \mathbb{R}$. Let $\mathcal{T}$ be a finite label set of treatment indicators (in this paper we consider only the binary case). Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \times \mathcal{T}$ such that $z = (\mathbf{x}, y, t) \in \mathcal{Z}$ means that $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$ and $t \in \mathcal{T}$. Let μ be an unknown probability distribution over $\mathcal{Z}$ such that $\forall z \in \mathcal{Z},\ \mu(z) > 0$. We assume that $\mathcal{X}$ is a compact convex space with respect to $\|\cdot\|_2$, thus there exists a constant $\mathbf{C}_x$ such that $\|\mathbf{x}\|_2 \leq \mathbf{C}_x$. Also, $|y| \leq \mathbf{C}_y$. A distance metric is a symmetric, positive definite function with two arguments from $\mathcal{X}$ such that $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. A distance metric must obey the triangle inequality. Let $\mathcal{S}_n$ denote a set of n observed units $\{s_1, ..., s_n\}$ drawn i.i.d. from μ such that $\forall i,\ s_i \in \mathcal{Z}$. We parameterize $\mathbf{d}$ with parameter $\mathcal{M}(\cdot)$, explicitly calling it $\mathbf{d}_{\mathcal{M}}$, and let $\mathcal{M}(\mathcal{S}_n)$ denote the parameter learned using MALTS methodology which is described in Section 4. For ease of notation, we will denote the observed sample of treated units as $\mathcal{S}_n^{(T)} := \{s_i^{(T)} = (\mathbf{x}_i, y_i, t_i) \mid s_i^{(T)} \in \mathcal{S}_n \text{ and } t_i = T\}$ and the observed sample of control units as $\mathcal{S}_n^{(C)} := \{s_i^{(C)} = (\mathbf{x}_i, y_i, t_i) \mid s_i^{(C)} \in \mathcal{S}_n \text{ and } t_i = C\}$. We assume no unobserved confounders and standard ignorability assumptions (Rubin 2005). For each individual unit $s_i = (\mathbf{x}_i, y_i, t_i) \in \mathcal{Z}$ we define its conditional average treatment effect (or individualized treatment effect) as $\tau(\mathbf{x}_i) = y^{(T)}(\mathbf{x}_i) - y^{(C)}(\mathbf{x}_i)$. For notational simplicity we sometimes refer $y^{(T)}(\mathbf{x}_i)$ as $y_i^{(T)}$ and $y^{(C)}(\mathbf{x}_i)$ as $y_i^{(C)}$ We use the $\widehat{\ }$(hat) notation to refer to estimated values.

Our goal is to minimize the expected loss between estimated treatment effects $\widehat{\tau}(\mathbf{x})$ and true treatment effects $\tau(\mathbf{x})$ across target population $\mu(z)$ (this can either be a finite or super-population).

Let the population expected loss be:

$$\mathbb{E}\left[\ell(\widehat{\tau}(\mathbf{x}), \tau(\mathbf{x}))\right] = \int \ell(\widehat{\tau}(\mathbf{x}), \tau(\mathbf{x}))d\mu = \int \ell(\hat{y}^{(T)}(\mathbf{x}) - \hat{y}^{(C)}(\mathbf{x}), y^{(T)}(\mathbf{x}) - y^{(C)}(\mathbf{x}))d\mu.$$

We use absolute loss, $\ell(a, b) = |a - b|$. For a finite random i.i.d. sample $\{s_i = (\mathbf{x}_i, y_i, t_i)\}_{i=1}^n$ from the distribution μ, we could estimate the sample average loss as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}^{(T)}(\mathbf{x}_i) - \hat{y}^{(C)}(\mathbf{x}_i), y^{(T)}(\mathbf{x}_i) - y^{(C)}(\mathbf{x}_i)),$$

where $y^{(T)}(\mathbf{x}_i)$ and $y^{(C)}(\mathbf{x}_i)$ are the counterfactual outcome values for the units in the sample $\{s_i = (\mathbf{x}_i, y_i, t_i) : i = 1, \ldots, n\}$. However, the difficulty in causal inference is that we only observe treatment outcomes $y^{(T)}(\mathbf{x}_i)$ or control outcomes $y^{(C)}(\mathbf{x}_i)$ for an individual i in the sample. Hence, we cannot directly calculate the treatment effect for any individual. For units in the treatment set we know $y^{(T)}(\mathbf{x}_i)$ and so we replace $\hat{y}^{(T)}(\mathbf{x}_i)$ by $y^{(T)}(\mathbf{x}_i)$, and analogously for units in the control set. Thus breaking the sum into treatment and control group:

$$\frac{1}{n_t}\sum_{i \in \text{treated}} \ell(y^{(T)}(\mathbf{x}_i) - \hat{y}^{(C)}(\mathbf{x}_i), y^{(T)}(\mathbf{x}_i) - y^{(C)}(\mathbf{x}_i)) + \frac{1}{n_c}\sum_{i \in \text{control}} \ell(\hat{y}^{(T)}(\mathbf{x}_i) - y^{(C)}(\mathbf{x}_i), y^{(T)}(\mathbf{x}_i) - y^{(C)}(\mathbf{x}_i)).$$

For a unit in the treatment set $s_i^{(T)}$, we use matching to estimate the control outcome $\hat{y}^{(C)}(\mathbf{x}_i)$ by an average of the control outcomes within its matched group that we can observe. Let us define the *matched group* MG under the distance metric $\mathbf{d}_{\mathcal{M}}$ parameterized by $\mathcal{M}$ for treated unit s_i in terms of the observed control units $\mathcal{S}_n^{(C)} = \{s_k^{(C)}\}_k$ indexed by k, which are the K-nearest-neighbors from set $\mathcal{S}_n$ under the distance metric $\mathbf{d}_{\mathcal{M}}$:

$$\text{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \{s_k^{(C)}\}_{k=1}^K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, C) := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(C)}} \mathbb{1}\Big(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\Big)\right] < K \right\}. \quad (1)$$

We allow reuse of units in multiple matched groups. Thus,

$$\hat{y}^{(C)}(\mathbf{x}_i) = \frac{1}{K} \sum_{k \in \text{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \{s_k^{(C)}\}_{k=1}^K)} y_k, \quad (2)$$

where K is the size of the matched group $\text{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \{s_k^{(C)}\}_{k=1}^K)$.

Our framework learns a distance metric from a separate training set of data (not the estimation data considered in the averages above), and we denote this training set by $\mathcal{S}_{tr}$. To learn $\mathbf{d}_{\mathcal{M}}$, we minimize the following:

$$\mathcal{M}(\mathcal{S}_{tr}) \in \arg\min_{\mathcal{M}} \left[\begin{array}{l} \sum_{s_i \in \mathcal{S}_{tr}^{(T)}} \left(y_i - \hat{y}^{(T)}(\mathbf{x}_i)\right)^2 \\ + \sum_{s_i \in \mathcal{S}_{tr}^{(C)}} \left(y_i - \hat{y}^{(C)}(\mathbf{x}_i)\right)^2 \end{array} \right],$$

where $\hat{y}^{(C)}(\mathbf{x}_i)$ is defined by Equations (1) and (2) including its dependence on the distance $\mathbf{d}_{\mathcal{M}}$, which is parameterized by $\mathcal{M}$, using the training data for creating matched groups. $\hat{y}^{T}(\mathbf{x}_i)$ is defined analogously.

Once $\mathcal{M}(\mathcal{S}_{tr})$ is learned from the training set, it is used for estimation on the estimation data.

3.1 Smooth Distance Metric and Treatment Effect Estimation

In this subsection, we discuss that if a distance metric is a smooth distance metric then we can estimate the individualized treatment effect using a finite sample with high probability. First, let us define a smooth distance metric.

Definition 1 ***(Smooth Distance Metric)*** $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ *is a smooth distance metric if there exists a monotonically increasing bounded function* $\delta_{\mathbf{d}}(\cdot)$ *with zero intercept, such that* $\forall z_i, z_l \in \mathcal{Z}$ *if* $t_i = t_l$ *and* $\mathbf{d}(x_i, x_l) \leq \epsilon$ *then* $|y_i - y_l| \leq \delta_{\mathbf{d}}(\epsilon)$.

In the following text, the function $1NN$ refers to the *1-nearest-neighbor* version of KNN which returns the nearest neighbor of the query point.

Theorem 1 *Given a smooth distance metric* $\mathbf{d}_{\mathcal{M}}$*, if we estimate individualized treatment effect* $\hat{\tau}(\cdot)$ *for any given* $z = (\mathbf{x}, y, t) \in \mathcal{Z}$ *by nearest neighbor matching on a finite sample*

$\mathcal{S}_n \overset{i.i.d}{\sim} \mu(\mathcal{Z}^n)$, *using distance metric* $\mathbf{d}_{\mathcal{M}}$, *then the estimated individualized treatment effect* $\hat{\tau}(\mathbf{x})$ *and the true individualized treatment effect* $\tau(\mathbf{x})$ *are farther than* ϵ *with probability less than* $\delta(\epsilon, \mathbf{d}_{\mathcal{M}}, n)$*:*

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \geq \epsilon\Big) \leq \delta(\epsilon, \mathbf{d}_{\mathcal{M}}, n).$$

Theorem 1 follows from Lemma 2 in the appendix which proves that we can estimate counterfactual outcomes y correctly with high probability using nearest neighbor matching under a smooth distance metric, and Lemma 3 in the appendix which proves that estimating counterfactual outcomes, y, correctly with high probability leads to estimating CATEs, τ, correctly with high probability. In Section 6, Figure 1(b) shows that as the size of the estimation set increases, the mean error-rate for predicting CATE using any smooth distance metric decreases. We also show that using the MALTS methodology described in Section 4, we achieve significantly lower error-rate than a predefined Mahalanobis distance metric.

4 Matching After Learning to Stretch (MALTS)

MALTS performs weighted nearest neighbors matching, where the weights for the nearest neighbors can be learned by minimizing the following objective:

$$\mathbf{W} \in \arg\min_{\widetilde{\mathbf{W}}} \left[\sum_{i \in \mathcal{S}_{tr}^{(T)}} \left\| y_i - \sum_{s_l \in \mathcal{S}_{tr}^{(T)}, i \neq l} \widetilde{W}_{i,l} y_l \right\| \right] + \left[\sum_{i \in \mathcal{S}_{tr}^{(C)}} \left\| y_i - \sum_{l \in \mathcal{S}_{tr}^{(C)}, i \neq l} \widetilde{W}_{i,l} y_l \right\| \right].$$

We let $\widetilde{W}_{i,l}$ be a function of $\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)$. For example, the $\widetilde{W}_{i,l}$ can encode whether l belongs to i's K-nearest neighbors. Alternatively they can encode soft KNN weights where $\widetilde{W}_{i,l} \propto e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}$.

As a reminder of our notation, we consider distance metric $\mathbf{d}_{\mathcal{M}}$ parameterized by a set of parameters $\mathcal{M}$. We use Euclidean distances for continuous covariates, namely distances of the form $\|\mathcal{M}\mathbf{x}_a - \mathcal{M}\mathbf{x}_b\|_2$ where $\mathcal{M}$ encodes the orientation of the data. Usually, $\mathcal{M}$ is hard-coded rather than learned; an example in the causal inference literature is the classical Mahalanobis distance ($\mathcal{M}$ is fixed as the inverse covariance matrix for the observed covariates). This approach has been demonstrated to perform well in settings where all covariates are observed and the inferential target is the average treatment effect (Stuart 2010a). We are interested instead in individualized treatment effects, and just as the choice of Euclidean norm in Mahalanobis distance matching depends on the estimand of interest, the stretch metric needs to be amended for this new estimand. We propose learning the parameters of distance metric, $\mathcal{M}$, directly from the observed data rather than setting it beforehand. The parameters of distance metric, $\mathcal{M}$, can be learned such that $\mathbf{W}$ minimizes the objective function on the training set.

We need to define "approximate closeness" differently for discrete covariates. If we use the same distance metric for both discrete and continuous data, then units that are close in continuous space might be arbitrarily far in discrete space or vice versa (e.g., a choice of either

Hamming distance or Euclidean distance would have this problem when used for both discrete and continuous covariates–Euclidean distance may not be defined for discrete covariates, whereas Hamming distance makes little sense for continuous covariates). Because of this, it is not natural to parameterize a single form of distance metric to enforce both exact matching on discrete data and almost-exact matching for continuous data. While Mahalanobis-distance-matching papers recommend converting unordered categorical variables to binary indicators (Stuart 2010b), this approach does not scale and in fact can introduce an overwhelming number of irrelevant covariates. Thus, mixed data poses a different set of challenges than either one alone, given the geometry of the space.

To accomodate continuous and discrete covariates, we parameterize our distance metric in terms of two components: one is a learned weighted Euclidean distance for continuous covariates while the other is a learned weighted Hamming distance for discrete covariates as in the FLAME and DAME algorithms (Wang et al. 2017, Dieng et al. 2019). These components are separately parameterized by matrices $\mathcal{M}_c$ and $\mathcal{M}_d$ respectively, $\mathcal{M} = [\mathcal{M}_c, \mathcal{M}_d]$. Let $a = (a_c, a_d)$ and $b = (b_c, b_d)$ be the covariates for two individuals split into continuous and discrete pairs respectively. The distance metric we propose is thus given by:

$$\text{distance}_{\mathcal{M}}(a, b) = d_{\mathcal{M}_c}(a_c, b_c) + d_{\mathcal{M}_d}(a_d, b_d), \text{ where}$$

$$d_{\mathcal{M}_c}(a_c, b_c) = \|\mathcal{M}_c a_c - \mathcal{M}_c b_c\|_2, \quad d_{\mathcal{M}_d}(a_d, b_d) = \sum_{j=0}^{|a_d|} \mathcal{M}_d^{(j,j)} \mathbb{1}[a_d^{(j)} \neq b_d^{(j)}],$$

and $\mathbb{1}[A]$ is the indicator that event A occurred. We thus perform learned Hamming distance matching on the discrete covariates and learned-Mahalanobis-distance matching for continuous covariates.

We separate the observed samples $\mathcal{S}_n$ into training set $\mathcal{S}_{tr}$ (not used for matching) and the estimation set $\mathcal{S}_{est}$. We thus learn $\mathcal{M}(\mathcal{S}_{tr})$ using the training sample $\mathcal{S}_{tr}$ such that

$$\mathcal{M}(\mathcal{S}_{tr}) \in \arg\min_{\mathcal{M}} \left(c\|\mathcal{M}\|_{\mathcal{F}} + \Delta_{\mathcal{S}_{tr}}^{(C)}(\mathcal{M}) + \Delta_{\mathcal{S}_{tr}}^{(T)}(\mathcal{M}) \right) \tag{3}$$

where, $\|\cdot\|_{\mathcal{F}}$ is Frobenius norm of the matrix,

$$\begin{aligned} \Delta_{\mathcal{S}_{tr}}^{(C)}(\mathcal{M}) :&= \sum_{s_i \in \mathcal{S}_{tr}^{(C)}} \left| y_i - \sum_{s_l \in \mathcal{S}_{tr}^{(C)}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}_{tr}^{(C)}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} y_l \right| \\ &= \sum_{s_i \in \mathcal{S}_{tr}^{(C)}} \left| \sum_{s_l \in \mathcal{S}_{tr}^{(C)}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}_{tr}^{(C)}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} (y_i - y_l) \right| \end{aligned} \tag{4}$$

$$\Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) := \sum_{s_i \in \mathcal{S}^{(T)}_{tr}} \left| y_i - \sum_{s_l \in \mathcal{S}^{(T)}_{tr}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(T)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} y_l \right| = \sum_{s_i \in \mathcal{S}^{(T)}_{tr}} \left| \sum_{s_l \in \mathcal{S}^{(T)}_{tr}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(T)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} (y_i - y_l) \right|. \tag{5}$$

For interpretability, we let $\mathcal{M}_c$ be a diagonal matrix, which allows stretches of the continuous covariates. This way, the magnitude of an entry in $\mathcal{M}_c$ or $\mathcal{M}_d$ provides the relative importance of the indicated covariate for the causal inference problem. We use python scipy library's implementation of COBYLA, a non-gradient optimization method, to learn $\mathcal{M}$ (Jones et al. 01 , Powell 1994).

We used the learned distance metric $\mathcal{M}(\mathcal{S}_{tr})$ to predict conditional average treatment effects (CATEs) for each unit in the estimation set, using its nearest neighbors from the same estimation set. For any given unit s in the estimation set, we construct a K-nearest neighbor matched group using control set $\mathcal{S}^{(C)}_{est}$ and using treatment set $\mathcal{S}^{(T)}_{est}$. Estimated CATE for a treated unit $s = (\mathbf{x}, y, t = T)$ is calculated via hard or soft KNN:

$$\hat{\tau}(\mathbf{x}) = y - \left(\sum_{s_i \in \mathcal{S}^{(C)}_{est}} \widetilde{W}_i y_i \right).$$

In this setting, choosing $\widetilde{W}_i$ to be proportional to $e^{\mathbf{d}_{\mathcal{M}}(\mathbf{x}, \mathbf{x}_i)}$ leads to robust and generatlizable CATE estimates via soft KNN (as shown in Theorem 2 and Theorem 3 below), while letting $\widetilde{W}_i$ be proportional to $\mathbb{1}\left[s_i \in \text{KNN}^{\mathcal{S}^{(C)}_{est}}_{\mathcal{M}(\mathcal{S}_{tr})} \right]$ produces reliable CATE estimates and interpretable matched groups.

5 Robustness and Generalization of MALTS

In this section we show that the MALTS framework estimates the correct distance metric and thus facilitates the correct estimates of CATEs. First, we define pairwise loss for s_i and s_l so that it is only finite for treatment-treatment or control-control matched pairs,

$$loss[\mathcal{M}, s_i, s_l] := \begin{cases} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)} |y_i - y_l| & \text{if } t_i = t_l \\ \infty & \text{otherwise.} \end{cases} \tag{6}$$

Further, we define an empirical average loss over finite sample $\mathcal{S}_n$ of size n as

$$L_{emp}(\mathcal{M}, \mathcal{S}_n) := \frac{1}{n^2} \sum_{(s_i, s_l) \in (\mathcal{S}_n \times \mathcal{S}_n)} loss[\mathcal{M}, s_i, s_l] \tag{7}$$

and define an average loss over population $\mathcal{Z}$ as

$$L_{pop}(\mathcal{M}, \mathcal{Z}) := \mathbb{E}_{z_i, z_l \overset{i.i.d}{\sim} \mu(\mathcal{Z})}\Big[loss[\mathcal{M}, z_i, z_l]\Big]. \tag{8}$$

Now, because the learned $\mathcal{M}(\mathcal{S}_{tr})$ on the set $\mathcal{S}_{tr}$ is the distance metric that minimizes the given objective function, we know that the following inequality is true, which states that the learned parameter has a lower training objective than that of the trivial parameter $\mathbf{0}$:

$$\Big(c\|\mathcal{M}(\mathcal{S}_{tr})\|_{\mathcal{F}} + \Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}(\mathcal{S}_{tr})) + \Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}(\mathcal{S}_{tr}))\Big) \leq \Big(c\|\mathbf{0}\|_{\mathcal{F}} + \Delta^{(C)}_{\mathcal{S}_{tr}}(\mathbf{0}) + \Delta^{(T)}_{\mathcal{S}_{tr}}(\mathbf{0})\Big) =: g_0. \tag{9}$$

Denoting the right hand side of the inequality by g_0 we note that we can limit our search space over distance metrics $\mathcal{M}$ that satisfy the following inequality:

$$\|\mathcal{M}\|_{\mathcal{F}} \leq \frac{g_0}{c}. \tag{10}$$

Thus, we observe that

$$\Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) \leq \sum_{s_i \in \mathcal{S}^{(C)}_{tr}} \sum_{s_l \in \mathcal{S}^{(C)}_{tr}} \left| \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(C)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} (y_i - y_l) \right| = \sum_{s_i \in \mathcal{S}^{(C)}_{tr}} \frac{\sum_{s_l \in \mathcal{S}^{(C)}_{tr}} loss[\mathcal{M}, s_i, s_l]}{\sum_{s_k \in \mathcal{S}^{(C)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}}.$$

We know that:

$$\forall i, k \ \mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)' \mathcal{M} (\mathbf{x}_i - \mathbf{x}_k) \leq \|\mathbf{x}_i - \mathbf{x}_k\|^2 \|\mathcal{M}\|_{\mathcal{F}} \leq \frac{g_0 \mathbf{C}_x^2}{c}.$$

Together, the two previous lines imply:

$$\Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) \leq \frac{1}{n \exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right)} \sum_{s_i \in \mathcal{S}^{(C)}_{tr}} \sum_{s_l \in \mathcal{S}^{(C)}_{tr}} loss[\mathcal{M}, s_i, s_l] = \frac{n L_{emp}(\mathcal{M}, \mathcal{S}^{(C)}_{tr})}{\exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right)}. \tag{11}$$

Similarly,

$$\Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) \leq \frac{n L_{emp}(\mathcal{M}, \mathcal{S}^{(T)}_{tr})}{\exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right)}. \tag{12}$$

Now, we define a few important concepts important for our results including covering number, smooth-distance-metric, robustness and generalizability.

Definition 2 *(**Covering Number**) Let $(\mathcal{U}, \rho)$ be a metric space. Consider a subset $\mathcal{V}$ of $\mathcal{U}$, then $\hat{\mathcal{V}} \subset \mathcal{V}$ is called a γ-cover of $\mathcal{V}$ if for any $v \in \mathcal{V}$, we can always find a $\hat{v} \in \hat{\mathcal{V}}$ such that $\rho(v, \hat{v}) \leq \gamma$. Further, the γ-covering-number of $\mathcal{V}$ under the distance metric ρ is defined by $\mathbf{N}(\gamma, \mathcal{V}, \rho) := \min\big\{|\hat{\mathcal{V}}| \ : \ \hat{\mathcal{V}}$ is a γ-cover of $\mathcal{V}\big\}$.*

Note that $\mathbf{N}(\gamma, \mathcal{V}, \rho)$ is finite if $\mathcal{U}$ is a compact.

Definition 3 *(**Robustness**) A learned distance metric $\mathcal{M}(\cdot)$ is $(K, \epsilon(\cdot))$-robust for a given K and $\epsilon(\cdot) : (\mathcal{Z} \times \mathcal{Z})^n \to \mathbb{R}$, if we can partition $\mathcal{X}$ into K disjoint sets $\{C_i\}_{i=1}^K$ such that for all samples $\mathcal{S}_{tr}$ and the corresponding pair set $\mathcal{S}_{tr}^2 := \mathcal{S}_{tr} \times \mathcal{S}_{tr}$ associated to the sample $\mathcal{S}_{tr}$, we have for any pair of training units $\big(s_1 = (\mathbf{x}_1, y_1, t_1), s_2 = (\mathbf{x}_2, y_2, t_2)\big) \in \mathcal{S}_{tr}^2$, and for any pair of units in the support $\big(z_1 = (\mathbf{x}_1', y_1', t_1'), z_2 = (\mathbf{x}_2', y_2', t_2')\big) \in \mathcal{Z}^2,\ \forall i, l \in \{1, ..., K\}$,*

$$\textit{if } \mathbf{x}_1, \mathbf{x}_1' \in C_i \textit{ and } \mathbf{x}_2, \mathbf{x}_2' \in C_l \textit{ such that } t_1 = t_1' = t_2 = t_2' \textit{ then}$$

$$\Big| \ loss[\mathcal{M}(\mathcal{S}_{tr}), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_{tr}), z_1, z_2] \ \Big| \leq \epsilon(\mathcal{S}_{tr}).$$

Intuitively, *robustness* means that for any possible units in the support, the loss is not far away from the loss of nearby units in training set, should some training units exist nearby. As the training procedure aims at minimizing the cumulative loss, we can safely say that a robust method will not perform poorly out of sample.

Definition 4 *(**Generalizability**) A learned distance metric $\mathcal{M}(\cdot)$ is said to generalize with respect to the given sample $\mathcal{S}_{tr}$ such that $|\mathcal{S}_{tr}| = n_{tr}$ if*

$$\lim_{n_{tr} \to \infty} \sum_{t \in \mathcal{T}} \Big| L_{pop}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{Z}^{(t)}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{S}_{tr}^{(t)}) \Big| = 0.$$

Given these definitions, we first show that the distance metric learned using MALTS is robust in Theorem 2 and we extend the argument to show that it is also generalizable in Theorem 3.

Theorem 2 *Given a fixed γ and smooth distance metric $\rho = \|\cdot\|_2$ with bounding function $\delta(\cdot)$, the distance metric $\mathcal{M}(\cdot)$ learned using* MALTS *is:*

$$\left(\mathbf{N}(\gamma, \mathcal{X}, \|\cdot\|_2), 2\mathbf{C}_y \Big| e^{\frac{4\mathbf{C}_x \gamma g_0}{c}} - 1 \Big| + 2\delta(\gamma) \right) \text{−robust.}$$

We have a detailed proof of Theorem 2 in the appendix.

Theorem 3 *The distance metric $\mathcal{M}(\cdot)$ learned using* MALTS *is generalizable.*

$$\lim_{n \to \infty} \left(\begin{array}{c} \Big| L_{pop}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{Z}^{(C)}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{S}_{tr}^{(C)}) \Big| \\ + \Big| L_{pop}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{Z}^{(T)}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{S}_{tr}^{(T)}) \Big| \end{array} \right) = 0.$$

Proof. By Theorem 2 we know that the distance metric $\mathcal{M}(\cdot)$ learned using MALTS is $(K, \epsilon(\cdot))$-robust where $K = \mathbf{N}(\gamma, \mathcal{X}, \|\cdot\|_2)$ and $\epsilon(\cdot) = 2\mathbf{C}_y \Big| e^{4\mathbf{C}_x \gamma g_0 / c} - 1 \Big| + 2\delta(\gamma)$. Using Lemma 1 (stated below), for any arbitrary $t' \in \mathcal{T}$ and $\mathcal{E} > 0$ we have

$$P_{\mathcal{S}_{tr}} \left(\begin{array}{c} \Big| L_{pop}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}), \mathcal{S}_{tr}^{(t')}) \Big| \\ \geq \epsilon(\mathcal{S}_{tr}^{(t')}) + 2B \sqrt{\frac{2K \ \ln(2) \ + \ 2 \ \ln(1/\mathcal{E})}{n_{tr}^{(t')}}} \end{array} \right) \leq \mathcal{E}.$$

Thus, for the sum over all possible $t' \in \mathcal{T} = \{T, C\}$ we have:

$$P_{\mathcal{S}_{tr}}\left(\begin{array}{c}\sum_{t'\in\mathcal{T}}\left|L_{pop}(\mathcal{M}(\mathcal{S}_{tr}),\mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}),\mathcal{S}_{tr}^{(t')})\right| \\ \geq \sum_{t'\in\mathcal{T}}\left(2\Big(\mathbf{C}_y\Big|e^{4\mathbf{C}_x\gamma g_0/c}-1\Big| + \delta(\gamma)\Big) + 2B\sqrt{\frac{2K\ \ln(2)\ +\ 2\ \ln(1/\mathcal{E})}{n_{tr}^{(t')}}}\right)\end{array}\right) \leq 2\mathcal{E}.$$

γ in Theorem 2 was arbitrary, allowing us to take it to 0 in such a way that K increases at a rate smaller than $n_{tr}^{(t')}$ increases. $\mathcal{E}$ was also set arbitrarily, allowing us to take it to 0 slowly enough such that as $n_{tr} \to \infty$, each of the $n_{tr}^{(t')} \to \infty$ we have:

$$\lim_{n_{tr}\to\infty}\left(\sum_{t'\in\mathcal{T}}\left|L_{pop}(\mathcal{M}(\mathcal{S}_{tr}),\mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_{tr}),\mathcal{S}_{tr}^{(t')})\right|\right) = 0.$$

Next, we state and prove Lemma 1 which we used to prove Theorem 3.

Lemma 1 *Given training sample $\mathcal{S}_{tr} \overset{i.i.d}{\sim} \mu(\mathcal{Z})$ where $n_{tr}^{(t')}$ is the number of units with $t_i = t'$ in $\mathcal{S}_{tr}$, and choosing $B > 0$ for which $loss[\cdot, z_i, z_l] \leq B\ \forall z_i, z_l \in \mathcal{Z}$ (B exists because $\mathcal{X}$ is compact and $\mathcal{Y}$ is bounded): if a learning algorithm $\mathcal{A}(\mathcal{S}_{tr})$ is $(K, \epsilon(\cdot))$-robust then for any $\mathcal{E} > 0$, with probability greater than or equal to $1 - \mathcal{E}$ we have*

$$\forall t' \in \mathcal{T},\ \left|L_{pop}(\mathcal{A}(\mathcal{S}_{tr}),\mathcal{Z}^{(t')}) - L_{emp}(\mathcal{A}(\mathcal{S}_{tr}),\mathcal{S}_{tr}^{(t')})\right| \leq \epsilon(\mathcal{S}_n^{(t')}) + 2B\sqrt{\frac{2K\ \ln(2)\ +\ 2\ \ln(1/\mathcal{E})}{n_{tr}^{(t')}}}\ .$$

We show a detailed proof of Lemma 1 in the appendix.

Now that we have theoretical proved the functionality of MALTS, we will next discuss and compare MALTS performance with other methods on different datasets.

6 Experiments

In this section, we discuss the the performance of MALTS on both synthetically generated datasets (continuous covariates and mixed covariates) and the canonical Lalonde dataset.

6.1 Continuous Covariates

We study MALTS' performance by analyzing trends in CATE estimation error rates and statistics on matched groups. The data generation processes (DGP) for experimentation includes quadratic treatment effect terms in addition to a linear treatment effect and linear baseline effect. We generate p covariates, with k of them contributing to the outcome, i.e., there are $p - k$ irrelevant covariates. Here is the first data generation process DGP-1 (with independent covariates):

$$t_i = 0: \quad \mathbf{x}_i \sim \mathcal{N}(\mu^{(C)}, \Sigma^{(C)}),\ \mu^{(C)} = \mathbb{1}_p,\ \Sigma^{(C)} = 0.5 \cdot \mathbb{I}_p,$$

$$t_i = 1: \quad \mathbf{x}_i \sim \mathcal{N}(\mu^{(T)}, \Sigma^{(T)}),\ \mu^{(T)} = 2\mathbb{1}_p,\ \Sigma^{(T)} = \mathbb{I}_p, \tag{DGP-1}$$
$$\forall j \in \{1, \ldots, p\}: \quad P(a_j = 1) = P(a_j = -1) = \frac{1}{2},\ \beta_j = a_j \cdot \frac{10}{2^j},\ \alpha_j \sim \mathcal{N}(1, 0.5)$$

where $\mathbb{1}_p$ is a vector of ones of length p and $\mathbb{I}_p$ is the $p \times p$ identity matrix. Data generation process DGP-2 (with correlated covariates) is:

$$\begin{aligned} &\mathbf{x}_i \sim \mathcal{N}(\mu, \Sigma), \mu = \mathbb{1}_p, \Sigma = (1-\rho)\mathbb{I}_p + \rho \mathbb{1}_p \mathbb{1}_p^T,\ \rho = 0.2, \\ &t_i \sim \text{Bernoulli}\left(0.5 \text{ expit}\left(\mathbf{x}_{i1} + \mathbf{x}_{i2}\right)\right) \\ \forall j \in \{1, \ldots, p\}: \quad &P(a_j = 1) = P(a_j = -1) = \frac{1}{2},\ \beta_j = a_j \cdot \frac{10}{2^j},\ \alpha_j \sim \mathcal{N}(1, 0.5) \end{aligned} \tag{DGP-2}$$

where $\text{expit}(a) = 1/(1+\exp(-a))$. For both DGP-1 and DGP-2, the outcome model is given by:

$$y_i = \sum_{j=1}^{k} \beta_j \mathbf{x}_{ij} + t_i \sum_{i=j}^{k} \alpha_j \mathbf{x}_{ij} + t_i \sum_{j,l=1, l>j}^{k} \mathbf{x}_{ij}\mathbf{x}_{il}.$$

Let us now discuss the results on these DGPs.

6.1.1 Variance within the Matched groups

We generated 10 independent training and testing sets where $p = k = 10$, $n_c = n_t = 1000$ and $n^{test} = 10000$. Recall that during the training phase, MALTS learns a distance metric that stretches the more relevant covariates while compressing the irrelevant covariates in order to better predict the outcome. Because the β's of the true model are exponentially decreasing in absolute value, MALTS should learn a distance metric where the stretch decreases in the order of the covariate indices. This ensures tighter matches on covariates that are more relevant to prediction of outcomes. A natural measure of covariate balance in the test set is the variability of a covariate within a matched group. Based on our data generation process, after running MALTS, we expect that the average variance is increasing in the order of covariate indices. For this collection of datasets, Figure 1 plots the average variances for each covariate. As expected, the variance is lower for the most important covariates and the variance stops increasing beyond the fourth covariate—that is, the matching mechanism does not necessarily distinguish between the importance of covariates above index four. As the contribution of these covariates to the outcome is less than 10% of the magnitude of the first covariate, this agrees with our intuition, based on knowledge of the data generation process.

6.1.2 Error-rate analysis

In this simulation, we compare MALTS with several other methods: BART (Chipman et al. 2010), CRF (Athey et al. 2015), difference of random forests, GenMatch (Diamond and Sekhon 2013) and Propensity Score Nearest Neighbor matching (Ross et al. 2015). We study two different settings, where in both settings, $n^{test} = 10000$ units. The setting are:

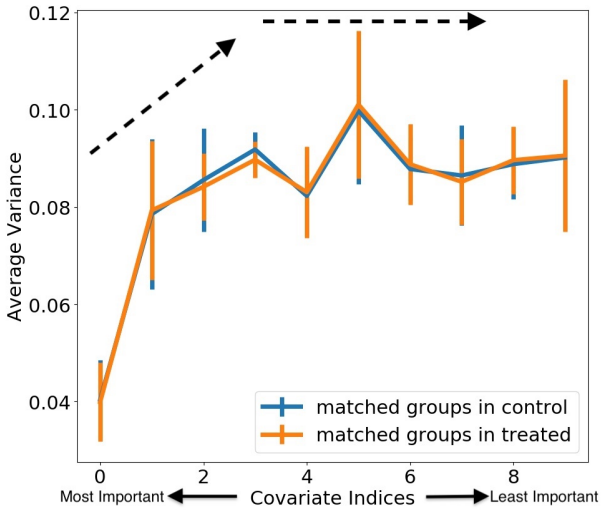


Figure 1: *Variance within matched groups tends to be smaller for important covariates*: Variance within matched groups for covariates during the testing phase of MALTS. Covariates are arranged in decreasing order of importance.

1) Uncorrelated covariates: Training set with $n_C = n_T = 1000$ units, $p = 10$ covariates observed, and $k = 8$ relevant covariates associated with the outcome.

2) Correlated covariates: Training set with $n = 2000$ units, $p = 18$ covariates observed, and $k = 8$ relevant covariates associated with the outcome.

Figures 2, 4 and 5 compares the CATE estimation error in both settings for different methods. We note that other matching methods are not designed for CATE estimation, hence they perform poorly in comparison to MALTS. MALTS is on par with modeling methods like causal forest and difference of random forests, which do not produce interpretable matches. MALTS does not outperform BART in our experiments, but recall that MALTS' distance metric was chosen to be inflexible (axis-aligned stretches) in order to maintain interpretability of the distance metric. We have similar findings when covariates are uncorrelated and when they are correlated.

Figure 3 is based on the uncorrelated simulated data and plots the reciprocal of the diameter of each matched group (where diameter is defined as the maximum distance of matched samples to the query sample in a matched group) versus the absolute CATE error. We note that tighter groups are of higher quality and lead to better estimation of CATEs. This suggests that we can threshold at a chosen diameter value to remove low quality matched groups or we can weight the matched group as a function of diameter for estimating a quantity of interest.

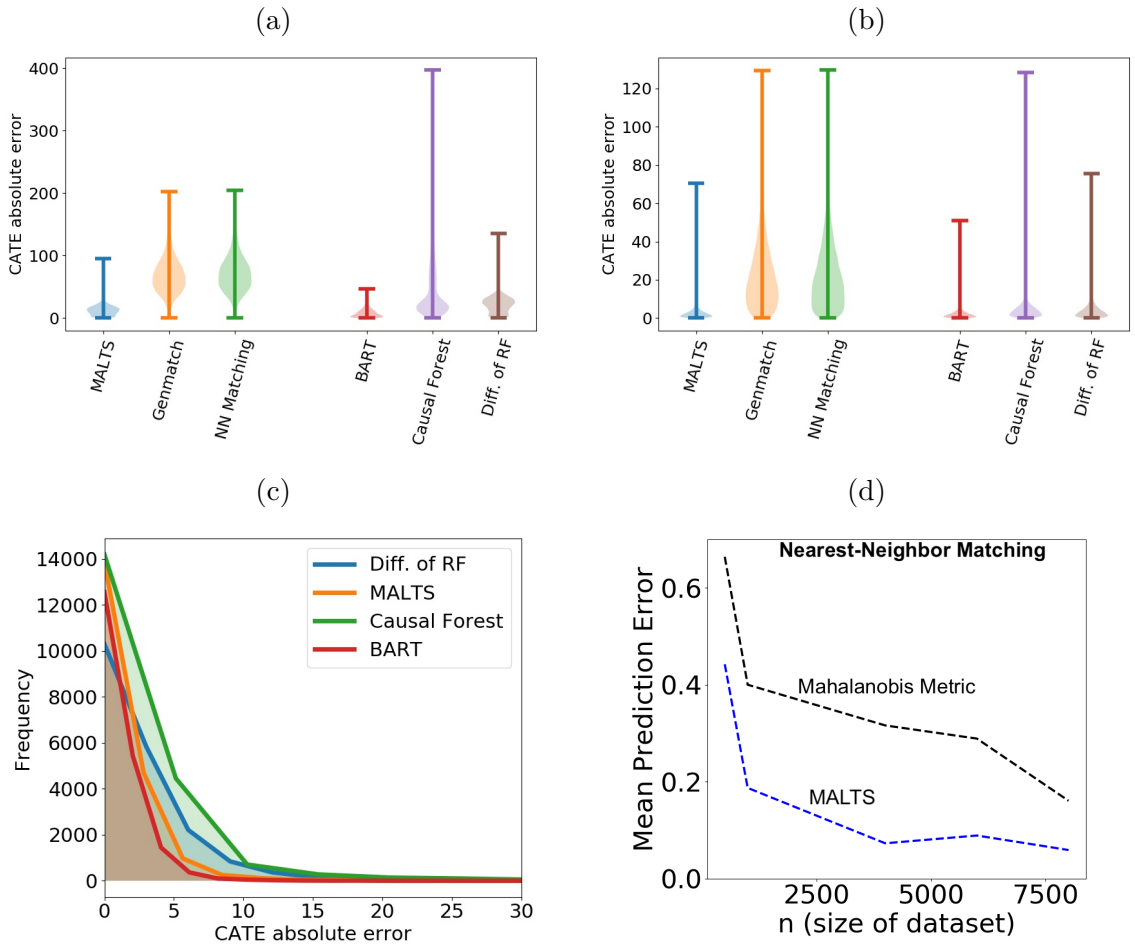


Figure 2: *MALTS performs well with respect to other methods.* Violin plots of CATE Absolute Error on the test set for several methods. MALTS performs well, despite being a matching method. (a) Uncorrelated covariates. (b) Correlated covariates with covariance matrix $\Sigma = (1-\rho)\mathbb{I}_p + \rho \mathbb{1}_p \mathbb{1}_p^T$ where $\rho = 0.2$. (c) Frequency plot of CATE absolute error for results shown in Figure 2(b). More detailed results on the distributions are in Figures 4 and 5. (d) *Amount of data to achieve low error for MALTS is small.* Mean Prediction Error for CATE decreases as the size of the estimation set increases. Also, MALTS' learned distance metric achieves a lower error rate compared to Mahalanobis smooth distance metric for nearest neighbor matching.

.

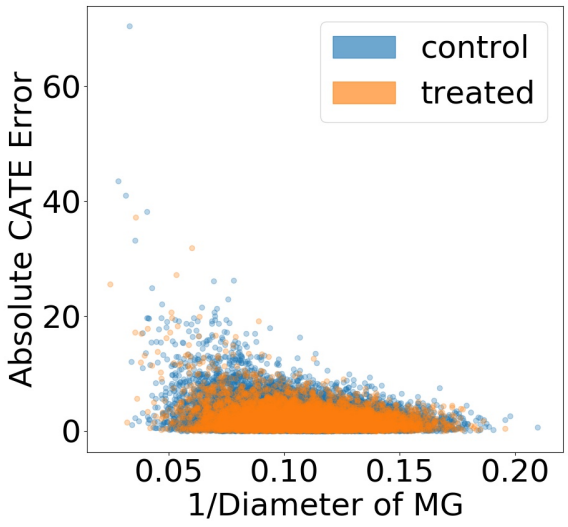


Figure 3: *Tighter matched groups are higher quality.* Scatter plot of CATE estimates for different matched groups versus the reciprocal of the diameter of the matched group. One may choose to prune matched groups based on diameter to remove less reliable large diameter groups (which are points on the left of the figure).

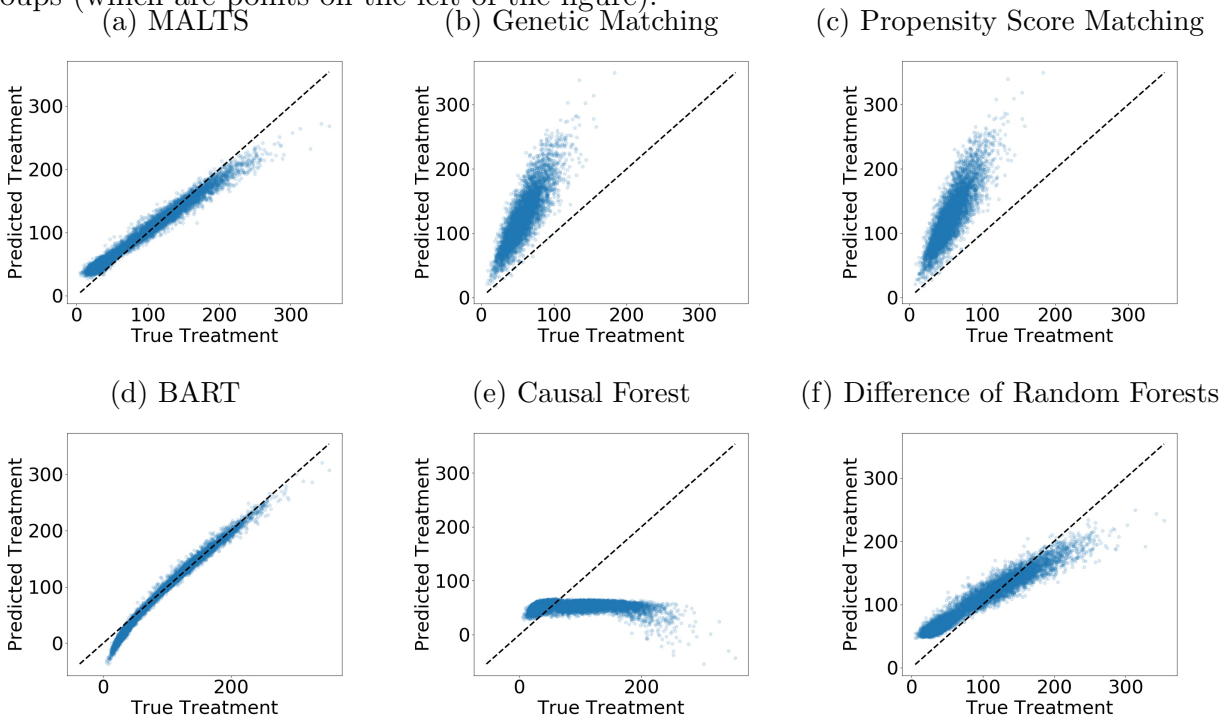


Figure 4: *Another view of performance of different methods, more detailed than Figure 2(a).* Scatter plots for true CATE vs predicted CATE for: (a) MALTS, (b) Genetic Matching, (c) Propensity Score Nearest Neighbor Matching, (d) difference of two BARTs, (e) Causal Forest and (f) difference of two random forests, where all the covariates are independent.

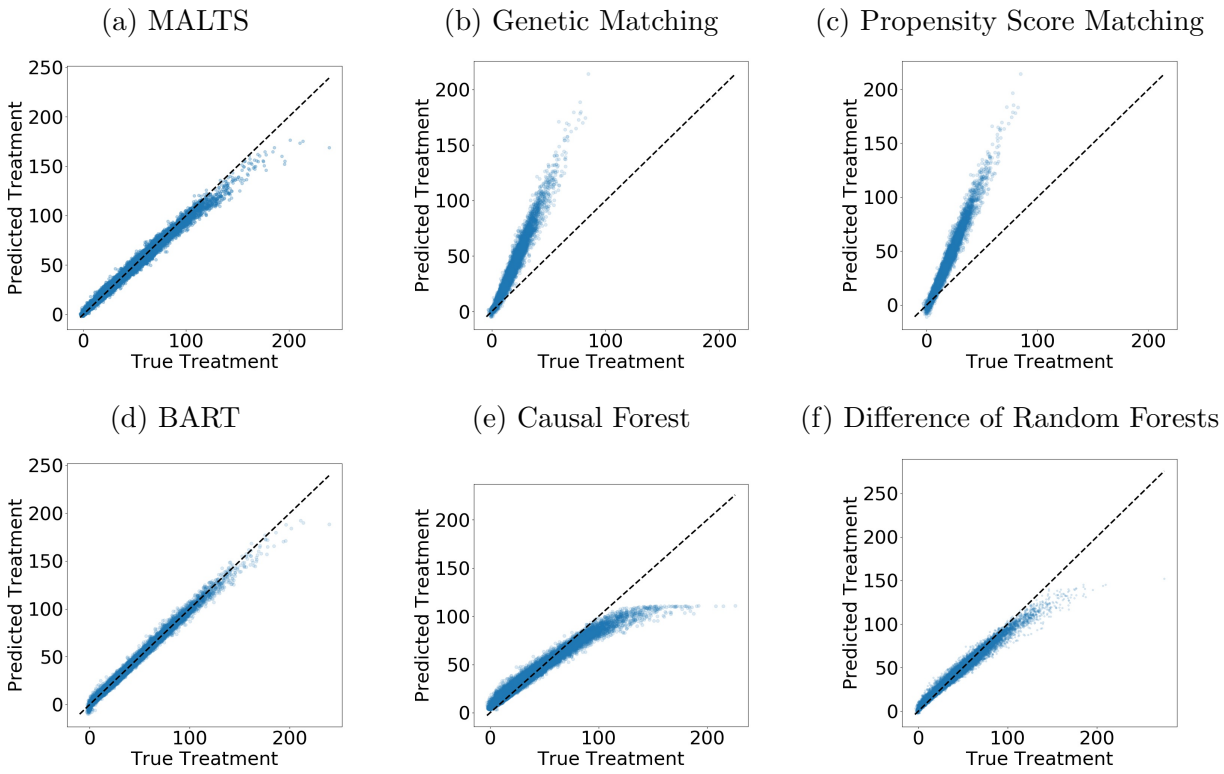


Figure 5: *Another view of performance of different methods, more detailed than Figure 2(b).* Scatter plots for true CATE vs predicted CATE for different estimation methods, where covariance matrix $\Sigma = (1-\rho)\mathbb{I}_p + \rho \mathbb{1}_p \mathbb{1}_p^T$ where $\rho = 0.2$.

6.2 Mixed Covariates

Let us denote $\mathbf{x}_{ci}$ to represent the continuous part of the covariates $\mathbf{x}_i$ for the i^{th} unit, and denote $\mathbf{x}_{di}$ to represent the discrete part. Here $\mathbf{x}_i$ is a p dimensional vector with p_c continuous covariates of which k_c are important in the determination of outcome y_i, and p_d discrete covariates of which k_d are important. We test MALTS on mixed data using the following DGP.

$$\begin{array}{ll} \textit{Continuous Variables:} & \mathbf{x}_{ci} \sim \mathcal{N}(\mu, \Sigma),\ \mu = \mathbb{1}_{p_c},\ \Sigma = 0.5 \cdot \mathbb{I}_{p_c} \\ \textit{Discrete Variable:} & \mathbf{x}_{dij} \overset{iid}{\sim} \textit{Bernoulli}(1/2), \\ \textit{Treatment Assignment:} & t_i \sim \textit{Bernoulli}(1/2). \end{array} \qquad \text{(DGP-3)}$$

We consider $p_c = 10$ continuous covariates where $k_c = 4$ are associated with the outcome and $p_d = 45$ discrete covariates where $k_d = 5$ are associated with the outcome. Letting $\tilde{\mathbf{x}}_i$ represent the 9 important covariates and letting $k = k_c + k_d = 9$ we generate the outcomes y_i according to the following quadratic model:

$$\textit{Outcome generation:} \quad y_i = \sum_{j=1}^{k} \beta_j \tilde{\mathbf{x}}_{ij} + t_i \sum_{j=1}^{k} \alpha_j \tilde{\mathbf{x}}_{ij} + t_i \sum_{j,l=1,l>j}^{k} \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{il}$$

$$\text{where} \quad P(a_j = 1) = P(a_j = -1) = \frac{1}{2}, \beta_j = \mathcal{N}(a_j 10, 1), \alpha_j \sim \mathcal{N}(1, 0.5).$$

Figure 6 provides a summary of CATE estimation error for different methods. MALTS continues to perform as well as some of the methods that directly model the outcome while outperforming the other (interpretable) matching methods. Figure 7 provides an in-depth view of these approaches. In particular, MALTS is better able to adapt to the unimportant discrete covariates than both CRF and the different of random forests.

6.3 Real Dataset: Lalonde

The Lalonde data pertain to the National Support Work Demonstration (NSW) temporary employment program and its effect on income level of the participants (LaLonde 1986). This dataset is frequently used as a benchmark for the performance of methods for observational causal inference. We employ the male sub-sample from the NSW in our analysis as well as the PSID-2 control sample of male household-heads under age 55 who did not classify themselves as retired in 1975 and who were not working when surveyed in the spring of 1976 (Dehejia and Wahba 1999). The outcome variable for both experimental and observational analyses is earnings in 1978 and the considered variables are age, education, whether a respondent is Black, is Hispanic, is married, has a degree, and their earnings in 1975. Previously, it has been demonstrated that almost any adjustment during the analysis of the experimental and observational variants of these data (both by modeling the outcome and by modeling the treatment variable) can lead to extreme bias in the estimate of average treatment effects. Tables 1 and 2 present the average treatment effect estimates based on MALTS, state-of-the-art modeling methods, and matching methods. MALTS (after appropriately thresholding or weighting for high-diameter matched groups according to the procedure described in Section 6.1.2) is able to achieve accurate ATE estimation on both

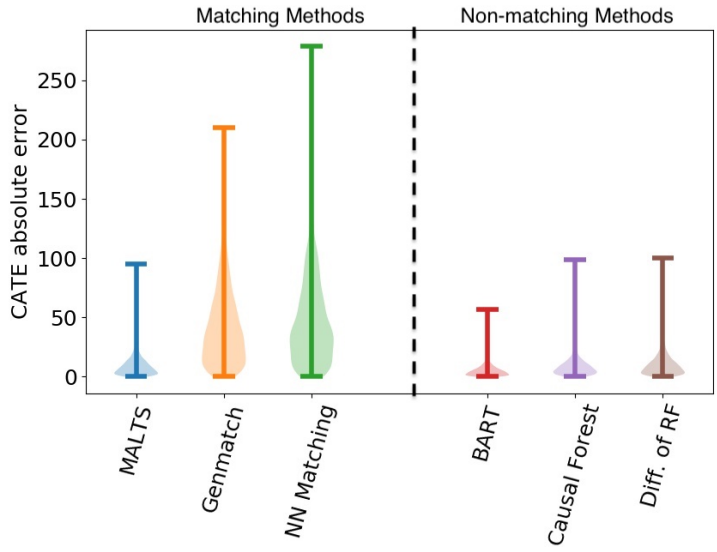


Figure 6: *MALTS performs well on mixed-data.* Violin plots of CATE absolute error on the test set for several methods on a dataset with mixed and independent covariates.

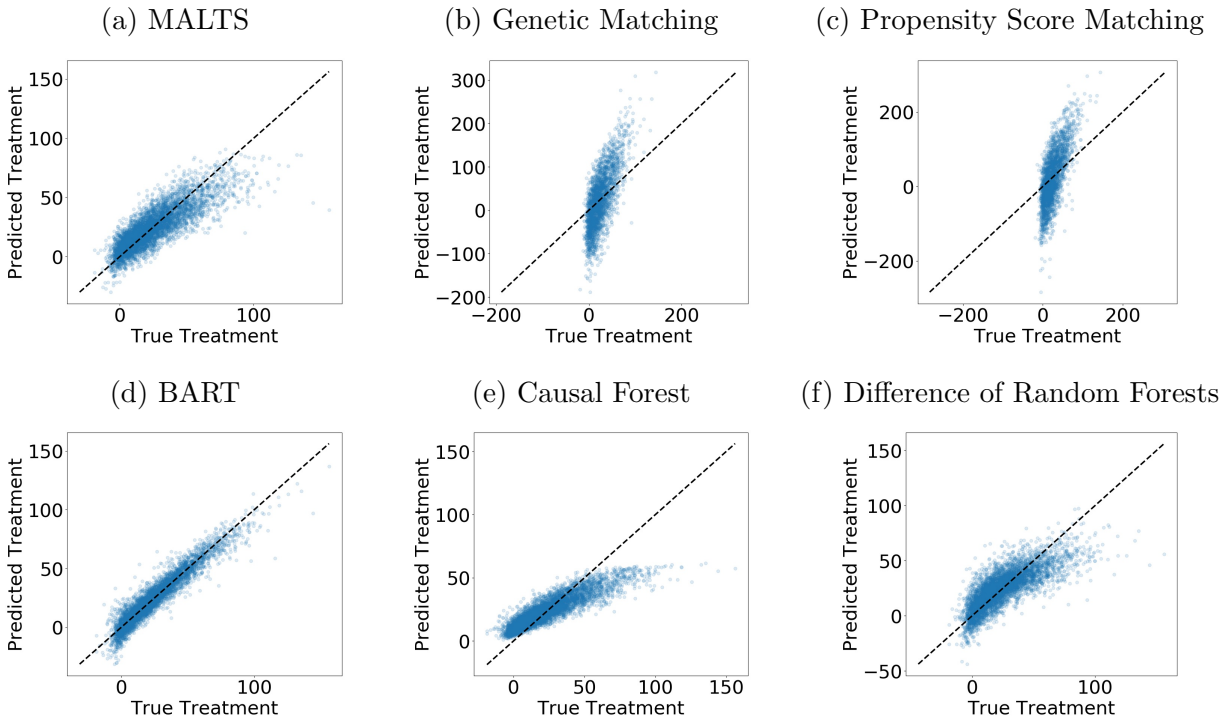


Figure 7: *Scatter plots for the true value of CATE vs predicted CATE for different methods* where all covariates are independent.

Table 1: *Predicted ATE for different methods on full Lalonde experimental dataset.* MALTS uses $w_i = e^{-0.4 * \text{diameter}(MG_i)}$ as the weight for matched group MG_i. Alternatively, we threshold to remove all matched groups with $\frac{1}{\text{diameter}(MG_i)} < 0.001$. In this setting, nearest neighbor matching is done with replacement. The "Number of units matched" column is filled only for matching methods.

Method	**ATE Estimate**	**Estimation Bias (%)**	**Number of units matched**
Truth	886	0	-
MALTS (Weighted)	885.82	-0.020	542
MALTS (Threshold)	890.27	0.48	489
CRF	781.20	-11.83	-
BART	836.11	-5.63	-
Genmatch	819.65	-7.48	594
Nearest Neighbor	825.27	-6.85	498

Table 2: *Predicted ATE for different methods on the PSID-2 control dataset and NSW treatment dataset.* MALTS uses $w_i = e^{-0.0015 * \text{diameter}(MG_i)}$ as the weight for matched-group MG_i for ATE estimation. Alternatively, we threshold to remove all matched groups with $1/\text{diameter}(MG_i) < 0.0008$. In this setting, nearest neighbor matching is done with replacement.

Method	**ATE Estimate**	**Estimation Bias (%)**	**Number of units matched**
Truth	886	0	-
MALTS (Weighted)	917.94	3.60	412
MALTS (Threshold)	921.44	4.00	277
CRF	-1359.66	-253.46	-
BART	-782.07	-188.27	-
Genmatch	-4007.69	-552.25	253
Nearest Neighbor	115.17	-87	372

experimental and observational datasets. The thresholding and weighting criteria for the observational Lalonde dataset were determined using the rule of thumb illustrated in Figure 8. We followed a similar procedure for the Lalonde experimental dataset to determine appropriate thresholding and weighting criteria.

To examine the interpretability of MALTS' matched groups, we present a sample of the matched groups from MALTS for the observational Lalonde dataset in Table 3. At the top of the table, we present the learned stretches for the distance metric ($\mathcal{M}$) and note that matching on age appears to be extremely important, followed by education. We present two individuals for whom we want to construct matched groups: Query 1 is an 23 year old individual with no income in 1975. We are able to construct a tight matched group for this individual (both in control and in treatment). In contrast, Query 2 is a 19-year-old high-income individual, which is an extremely unlikely scenario, leading to a matched group with a very large diameter, which should probably not be used during analysis.

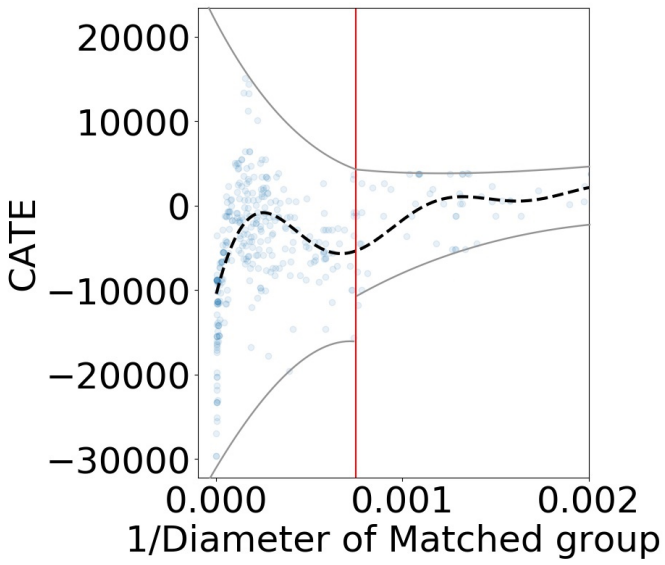


Figure 8: (*Rule of thumb for deciding a threshold*) Scatter Plot showing variation of CATEs with $\frac{1}{\text{diameter}(MG_i)}$ (surrogate for matched-groups' quality) for PSID-2 dataset. The red line shows the point where we threshold. The black trend curve is fitted using support vector regression. The grey line (envelope) is generated using a degree 2 polynomial.

Table 3: *Example of Matched-groups on Lalonde Experimental treatment and Observational control Datasets for two example query points drawn from the same datasets.* Query 1 represents a high quality (low diameter) matched group while Query 2 represents a poor quality (high diameter) matched group that could be discarded during analysis.

	Age	**Education**	**Black**	**Hispanic**	**Married**	**Nodegree**	**Income 1975 (re75)**
$\mathcal{M}$ (Diagonal Stretch Matrix)	2.745	1.61	0.331	0.389	0.206	0.434	0.164

	Age	**Education**	**Black**	**Hispanic**	**Married**	**No degree**	**Income in 1975 (re75)**	**Income in 1978 (re78)**	**T**
Query-1	23	12	1	0	0	0	0	4728.73	0
	22	12	1	0	1	0	0	664.98	0
	22	12	1	0	1	0	0	0	0
	24	12	1	0	0	0	0	10344.09	0
	25	12	1	0	0	0	0	0	0
	27	12	1	0	1	0	0	11821.81	0
	23	12	1	0	0	0	0	0	1
	23	12	1	0	0	0	0	4843.18	1
	22	12	1	0	0	0	0	18678.08	1
	25	12	1	0	0	0	0	2348.97	1
	25	12	1	0	0	0	0	0	1
Mean	*23.73*	*12*	*1*	*0*	*0.27*	*0*	*0*	*4870.11*	
Query-2	19	10	1	0	0	1	1374.59	3228.5	1
	24	13	0	0	1	0	1790.32	449.23	0
	26	12	1	0	1	0	1074.19	8866.36	0
	49	8	0	1	0	1	1074.19	0	0
	51	8	0	0	1	1	1546.84	0	0
	52	6	1	0	1	1	1246.06	1477.73	0
	18	10	1	0	0	1	1371.02	12064.41	1
	18	9	1	0	0	1	1285.33	15369.48	1
	17	9	1	0	0	1	1341.65	1374.07	1
	21	11	0	1	0	1	1395.39	0	1
	26	11	1	0	1	1	1392.85	1460.36	1
Mean	*30.2*	*9.7*	*0.6*	*0.2*	*0.5*	*0.8*	*1351.79*	*4106.16*	

7 Conclusion and Discussion

This paper introduced the MALTS algorithm, which learns a distance-metric on the covariate space for use with matching. The learned metric stretches important covariates and compresses irrelevant covariates for outcome prediction in order to produce high-quality matches. Unlike black-box machine learning methods, MALTS produces interpretable matched groups and returns the stretch matrix on covariates for counterfactual prediction. A natural extension that we are pursuing is to use neural networks or support vector machines to learn a flexible distance metric in a latent space, thus allowing us to match on medical records, images, and text documents. This will allow us to incorporate complex data structures by introducing a flexible learning framework (e.g., neural networks) for coding the data. That is, we can redefine the distance metric via

$$\begin{aligned} \text{distance}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_j) &= \langle \phi_{\mathcal{M}}(\mathbf{x}_i), \phi_{\mathcal{M}}(\mathbf{x}_j) \rangle \quad \text{or} \\ \text{distance}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_j) &= \left(\phi_{\mathcal{M}}(\mathbf{x}_i) - \phi_{\mathcal{M}}(\mathbf{x}_j)\right)^2, \end{aligned}$$

where ϕ is a summary of relevant data features learned using a complex modeling framework. As deep neural networks mainly show improvements over other methods for problems that do not have natural data representations (computer vision, speech, etc.), we conjecture that the stretch/almost-exact match combination should suffice for most datasets. The MALTS framework can be further extended to deal with missing covariates, and can be adapted to instrumental variables, which is an ongoing effort.

References

Athey, S., Eckles, D., and Imbens, G. W. (2015). Exact p-values for network interference. Technical report, National Bureau of Economic Research.

Bellet, A. and Habrard, A. (2015). Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. (2016). Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, pages 266–298.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95(3):932–945.

Dieng, A., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2019). Interpretable almost-exact matching for causal inference. *Proceedings of Machine Learning Research (Proceedings of AISTATS)*, 89:2445.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

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.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2018-11-10].

Kallus, N. (2017). A Framework for Optimal Matching for Causal Inference. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 372–381, Fort Lauderdale, FL, USA.

Keele, L. and Zubizarreta, J. R. (2014). Optimal multilevel matching in clustered observational studies: A case study of the school voucher system in chile. *arXiv preprint arXiv:1409.8597*.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76(4):604–620.

Noor-E-Alam, M. and Rudin, C. (2015a). Robust nonparametric testing for causal inference in natural experiments. Working paper.

Noor-E-Alam, M. and Rudin, C. (2015b). Robust testing for causal inference in natural experiments.

Parikh, H., Rudin, C., and Volfovsky, A. (2019). An application of matching after learning to stretch (MALTS) to the ACIC 2018 causal inference challenge data. *Observational Studies*, 5:118–130.

Powell, M. J. D. (1994). *A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation*, pages 51–67. Springer Netherlands, Dordrecht.

Resa, M. and Zubizarreta, J. R. (2016). Evaluation of subset matching methods and forms of covariate balance. *Statistics in Medicine*.

Rosenbaum, P. R. (2016). Imposing minimax and quantile constraints on optimal matching in observational studies. *Journal of Computational and Graphical Statistics*, (just-accepted).

Ross, M. E., Kreider, A. R., Huang, Y.-S., Matone, M., Rubin, D. M., and Localio, A. R. (2015). Propensity score methods for analyzing observational data like randomized experiments: challenges and solutions for rare outcomes and exposures. *American journal of epidemiology*, 181(12):989–995.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, pages 109–120.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331.

Stuart, E. A. (2010a). Matching methods for causal inference: A review and a look forward. *Statist. Sci.*, 25(1):1–21.

Stuart, E. A. (2010b). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

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*.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231.

Appendix A

In this section we provide proofs for theorems and lemmas discussed in section 5.

Proof (Theorem 2). Given $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \times \mathcal{T}$, we consider the following definition of a minimum sized γ-cover $\hat{\mathcal{V}}$ of the set $\mathcal{X}$ under the distance metrix $\|\cdot\|_2$: Partition the set into K disjoint subsets $\{C_i\}_{i=1}^K$ such that K is the γ-covering-number of $\mathcal{X}$ under $\|\cdot\|_2$ (which is exactly equal to $|\hat{\mathcal{V}}|$) where each C_i is the γ-neighborhood of each $\hat{v}_i \in \hat{\mathcal{V}}$ and each C_i contains at least one control and one treated sample. Note that if such a cover exists, then since $\mathcal{X}$ is a compact convex set, K is finite.

We further assume that distance metric $\|\cdot\|_2$ is a smooth distance metric with bounding function $\delta(\cdot)$. This implies that $\delta(\cdot)$ is a monotonically increasing zero-intercept function such that $\forall z_1, z_2 \in \mathcal{Z}$ if $t_1 = t_2$ and $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon$ then $|y_1 - y_2| \leq \delta(\epsilon)$. This further implies that for any z_1 and z_2 such that $x_1, x_2 \in C_i$ and $t_1 = t_2$ then, since they are both in the same ball of radius γ, we have $|y_1 - y_2| \leq \delta(\gamma)$.

For some $s_1 = (\mathbf{x}_1, y_1, t_1)$ and $s_2 = (\mathbf{x}_2, y_2, t_2)$ in the training set $\mathcal{S}_n$ and $z_1 = (\mathbf{x}_1', y_1', t_1')$ and $z_2 = (\mathbf{x}_2', y_2', t_2')$ in $\mathcal{Z}$ such that $\mathbf{x}_1, \mathbf{x}_1' \in C_i$, $\mathbf{x}_2, \mathbf{x}_2' \in C_l$, and $t_1 = t_1' = t_2 = t_2'$, then we try to bound the following quantity:

$$\left| loss[\mathcal{M}(\mathcal{S}_n), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_n), z_1, z_2] \right| = \left| |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1, \mathbf{x}_2)}(y_1 - y_2)| - |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1', \mathbf{x}_2')}(y_1' - y_2')| \right|.$$

From the reverse triangle inequality we know

$$\left| |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1, \mathbf{x}_2)}(y_1 - y_2)| - |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1', \mathbf{x}_2')}(y_1' - y_2')| \right|$$

$$
\begin{aligned}
&\leq \left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)}(y_1 - y_2) - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y'_1 - y'_2) \right| \\
&= \Big| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} y_1 - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} y_1 + e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} y_1 - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} y'_1 \\
&\quad - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} y_2 + e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} y_2 - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} y_2 + e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} y'_2 \Big| \\
&= \Bigg| y_1 \Big(e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \Big) + e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y_1 - y'_1) \\
&\quad - y_2 \Big(e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \Big) - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y_2 - y'_2) \Bigg|,
\end{aligned}
$$

and applying the triangle inequality,

$$
\begin{aligned}
&\leq \left| y_1 \Big(e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \Big) \right| + \left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y_1 - y'_1) \right| \\
&\quad + \left| y_2 \Big(e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \Big) \right| + \left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y_2 - y'_2) \right|.
\end{aligned}
$$

For any $y \in \mathcal{Y}$, we know that $|y| \leq \mathbf{C}_y$. Thus,

$$
\begin{aligned}
&\Big| |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)}(y_1 - y_2)| - |e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)}(y'_1 - y'_2)| \Big| \\
&\quad \leq 2\mathbf{C}_y \left(\left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \right| \right) + |y_1 - y'_1| + |y_2 - y'_2|.
\end{aligned}
$$

By the smoothness of distance metric $\|\cdot\|_2$ and the fact that the two points are in the same γ-sized ball, we know that $|y_1 - y'_1| + |y_2 - y'_2| \leq 2\delta(\gamma)$. Hence,

$$
\begin{aligned}
&\Big| loss[\mathcal{M}(\mathcal{S}_n), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_n), z_1, z_2] \Big| \\
&\quad \leq 2 \left(\mathbf{C}_y \left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \right| + \delta(\gamma) \right).
\end{aligned}
$$

If we multiply the right-hand-side of the inequality with a number greater than 1, then the inequality will not change. Hence,

$$
\begin{aligned}
&2 \left(\mathbf{C}_y \left| e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - e^{-\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}'_1,\mathbf{x}'_2)} \right| + \delta(\gamma) \right) \\
&\leq 2 \left(\mathbf{C}_y \left| e^{\mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(z_1,z_2) - \mathbf{d}_{\mathcal{M}(\mathcal{S}_n)}(\mathbf{x}_1,\mathbf{x}_2)} - 1 \right| + \delta(\gamma) \right) \\
&= 2 \left(\mathbf{C}_y \left| e^{(\mathbf{x}'_1 - \mathbf{x}'_2)^T \mathcal{M}(\mathcal{S}_n)(\mathbf{x}'_1 - \mathbf{x}'_2) - (\mathbf{x}_1 - \mathbf{x}_2)^T \mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1 - \mathbf{x}_2)} - 1 \right| + \delta(\gamma) \right)
\end{aligned}
$$

$$
\begin{aligned}
&= 2\Bigg(\mathbf{C}_y\Big|exp\Bigg((\mathbf{x}_1'-\mathbf{x}_2')^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_2')-(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_2')\\
&\quad+(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_2')-(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1-\mathbf{x}_2)\Bigg)-1\Big|+\delta(\gamma)\Bigg)\Bigg)\\
&= 2\Bigg(\mathbf{C}_y\Big|exp\Bigg((\mathbf{x}_1'-\mathbf{x}_2')^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_2'-\mathbf{x}_1+\mathbf{x}_2)\\
&\quad+(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_2'-\mathbf{x}_1+\mathbf{x}_2)\Bigg)-1\Big|+\delta(\gamma)\Bigg)\Bigg)\\
&= 2\Bigg(\mathbf{C}_y\Big|exp\Bigg((\mathbf{x}_1'-\mathbf{x}_2')^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_1)+(\mathbf{x}_1'-\mathbf{x}_2')^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_2-\mathbf{x}_2')\\
&\quad+(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_1'-\mathbf{x}_1)+(\mathbf{x}_1-\mathbf{x}_2)^T\mathcal{M}(\mathcal{S}_n)(\mathbf{x}_2-\mathbf{x}_2')\Bigg)-1\Big|+\delta(\gamma)\Bigg)\Bigg)\\
&\leq 2\Bigg(\mathbf{C}_y\Bigg|\exp\Big(\|\mathbf{x}_1-\mathbf{x}_2\|_2\|\mathcal{M}(\mathcal{S}_n)\|_{\mathcal{F}}\|\mathbf{x}_1-\mathbf{x}_1'\|_2+\|\mathbf{x}_1-\mathbf{x}_2\|_2\|\mathcal{M}(\mathcal{S}_n)\|_{\mathcal{F}}\|\mathbf{x}_2-\mathbf{x}_2'\|_2\\
&\quad+\|\mathbf{x}_1'-\mathbf{x}_2'\|_2\|\mathcal{M}(\mathcal{S}_n)\|_{\mathcal{F}}\|\mathbf{x}_1-\mathbf{x}_1'\|_2+\|\mathbf{x}_1'-\mathbf{x}_2'\|_2\|\mathcal{M}(\mathcal{S}_n)\|_{\mathcal{F}}\|\mathbf{x}_2-\mathbf{x}_2'\|_2\Big)-1\Bigg|+\delta(\gamma)\Bigg)\\
&\leq 2\Bigg(\mathbf{C}_y\Big|e^{4\mathbf{C}_x\gamma\|\mathcal{M}(\mathcal{S}_n)\|_{\mathcal{F}}}-1\Big|+\delta(\gamma)\Bigg)=2\mathbf{C}_y\Bigg(\Big|e^{4\mathbf{C}_x\gamma g_0/c}-1\Big|\Bigg)+2\delta(\gamma).
\end{aligned}
$$

Hence, we conclude that our fixed γ, the distance metric learned using MALTS algorithm is robust by the definition of robustness.

Proof (Lemma 1). If $(D_1,\ldots,D_K)$ is the multinomially distributed random vector with parameters d and $p_1,\ldots,p_K$ then, by the Bretagnolle-Huber-Carol inequality, $P(\sum_{i=1}^K\left|\frac{D_i}{d}-p_i\right|\geq\lambda)\leq 2^K e^{-\frac{d\lambda^2}{2}}$. Thus, for our case, we can consider N_i corresponding to the set of indices of units in sample $\mathcal{S}_n^{(t')}$ such that their x's are contained in the partition $\mathbf{C}_i$ as in Theorem 2. Hence, by the Bretagnolle-Huber-Carol inequality (Bellet and Habrard 2015), we know that

$$
P\Bigg(\sum_{i=1}^K\Big|\frac{|N_i|}{n^{(t')}}-\mu(\mathbf{C}_i)\Big|\geq\sqrt{\frac{2K\ \ln(2)\ +\ 2\ \ln(1/\mathcal{E})}{n^{(t')}}}\Bigg)\leq\mathcal{E}\ .
$$

Now, for some arbitrary $t'\in\mathcal{T}$ let us consider $\Big|L_{pop}(\mathcal{M}(\mathcal{S}_n),\mathcal{Z}^{(t')})-L_{emp}(\mathcal{M}(\mathcal{S}_n),\mathcal{S}_n^{(t')})\Big|$. We know that

$$
\Big|L_{pop}(\mathcal{M}(\mathcal{S}_n),\mathcal{Z}^{(t')})-L_{emp}(\mathcal{M}(\mathcal{S}_n),\mathcal{S}_n^{(t')})\Big|
$$

$$
\begin{aligned}
= & \left| \sum_{i,j=1}^{K} \Big(\mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1 = (\mathbf{x}_1', y_1', t_1'), z_2 = (\mathbf{x}_2', y_2', t_2')) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \mu(\mathbf{C}_i)\mu(\mathbf{C}_j) \Big) \right. \\
& \left. - \frac{1}{(n^{(t')})^2} \sum_{s_1,s_2 \in \mathcal{S}_n^{(t')}} loss(\mathcal{M}(\mathcal{S}_n), s_1, s_2) \right| \\
= & \left| \sum_{i,j=1}^{K} (\mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \mu(\mathbf{C}_i)\mu(\mathbf{C}_j)) \right. \\
& - \sum_{i,j=1}^{K} \left(\mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \mu(\mathbf{C}_i) \ \frac{|N_j|}{n^{(t')}} \right) \\
& + \sum_{i,j=1}^{K} \left(\mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \mu(\mathbf{C}_i) \ \frac{|N_j|}{n^{(t')}} \right) \\
& + \sum_{i,j=1}^{K} \mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \frac{|N_i|}{n^{(t')}} \frac{|N_j|}{n^{(t')}} \\
& - \sum_{i,j=1}^{K} \mathbb{E}_{\mathbf{x}_1',\mathbf{x}_2'}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \frac{|N_i|}{n^{(t')}} \frac{|N_j|}{n^{(t')}} \\
& \left. - \frac{1}{(n^{(t')})^2} \sum_{s_1,s_2 \in \mathcal{S}_n^{(t')}} loss(\mathcal{M}(\mathcal{S}_n), s_1, s_2) \right| \\
\leq & \left| \sum_{i,j=1}^{K} \mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \mu(\mathbf{C}_i) \Big(\mu(\mathbf{C}_j) - \frac{|N_j|}{n^{(t')}} \Big) \right| \\
& + \left| \sum_{i,j=1}^{K} \mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \frac{|N_j|}{n^{(t')}} \Big(\mu(\mathbf{C}_i) - \frac{|N_i|}{n^{(t')}} \Big) \right| \\
& + \left| \sum_{i,j=1}^{K} \mathbb{E}_{z_1,z_2}[loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2) \mid \mathbf{x}_1' \in \mathbf{C}_i, \mathbf{x}_2' \in \mathbf{C}_j] \ \frac{|N_i|}{n^{(t')}} \frac{|N_j|}{n^{(t')}} \right. \\
& \left. - \frac{1}{(n^{(t')})^2} \sum_{s_1,s_2 \in \mathcal{S}_n^{(t')}} loss(\mathcal{M}(\mathcal{S}_n), s_1, s_2) \right| \\
\leq & \ 2B \sum_{i=1}^{K} \left| \frac{|N_i|}{n^{(t')}} - \mu(\mathbf{C}_i) \right| + \epsilon(\mathcal{S}_n^{(t')}) \text{ where } B \text{ is } \max_{z_1,z_2} loss(\mathcal{M}(\mathcal{S}_n), z_1, z_2).
\end{aligned}
$$

Hence, we can conclude for all $t' \in \mathcal{T}$ we have

$$
P_{\mathcal{S}_n} \left(\left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| \geq \epsilon(\mathcal{S}_n^{(t')}) + 2B \sqrt{\frac{2K \ \ln(2) \ + \ 2 \ \ln(1/\mathcal{E})}{n^{(t')}}} \right) \leq \mathcal{E} \, .
$$

Lemma 2 *Given a smooth distance metric $\mathcal{M}$ and treatment choice variable $t' \in \mathcal{T}$, if we estimate the counterfactual $\hat{y}^{(t')}(\mathbf{x})$ for any given $z = (\mathbf{x}, y, t) \in \mathcal{Z}$ by nearest neighbor matching on a finite sample $\mathcal{S}_n \overset{i.i.d}{\sim} \mu(\mathcal{Z}^n)$ using distance metric $\mathcal{M}$, then the estimated counterfactual $\hat{y}^{(t')}(\mathbf{x})$ and the true counterfactual $y^{(t')}(\mathbf{x})$ are farther than ϵ with probability less than $\delta(\epsilon, \mathcal{M}, n)$,*

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})| \geq \epsilon\Big) \leq \delta(\epsilon, \mathcal{M}, n).$$

Proof (Lemma 2). Let $\mathcal{B}_{\mathcal{M}}(\mathbf{x}, r)$ be an open ball of radius $r > 0$ under distance metric $\mathcal{M}$, centered around point a fixed point $\mathbf{x} \in \mathcal{X}$. We know that there is a nonzero probability mass around any point $\mathbf{x} \in \mathcal{X}$,

$$\forall r > 0, P_{X \sim \mu(\mathcal{X})}(X \in \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r)) > 0. \tag{13}$$

As $\mathcal{S}_n = \{Z_1 = (X_1, Y_1, T_1), \ldots, Z_n = (X_n, Y_n, T_n)\} \overset{i.i.d}{\sim} \mu(\mathcal{Z})$, the probability that no unit Z_i with $T_i = t'$ from a n-sized random sample $\mathcal{S}_n = \{Z_i\}_{i=1}^n$ lies within the r-neighborhood of a given unit $z = (\mathbf{x}, y, t) \in \mathcal{Z}$ is

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Bigg(\bigwedge_{Z_i \in \mathcal{S}_n}\Big(X_i \notin \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r) \wedge T_i = t'\Big)\Bigg) \leq P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Bigg(\bigwedge_{Z_i \in \mathcal{S}_n}\Big(X_i \notin \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r)\Big)\Bigg) \tag{14}$$

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Bigg(\bigwedge_{Z_i \in \mathcal{S}_n}\Big(X_i \notin \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r)\Big)\Bigg) = \Big(1 - P_{X \sim \mu(\mathcal{X})}(X \in \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r))\Big)^n. \tag{15}$$

From Equation 14, we can deduce that the probability that every unit with $T_i = t'$ in randomly drawn sample $\mathcal{S}_n$ is at least at a distance r from a given $z = (\mathbf{x}, y, t)$ is

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(\min_{\substack{Z_i \in \mathcal{S}_n \\ T_i = t'}} \mathbf{d}_{\mathcal{M}}(X_i, \mathbf{x}) \geq r\Big) \leq \Big(1 - P_{X \sim \mu(\mathcal{X})}(X \in \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r))\Big)^n. \tag{16}$$

We infer from Equation 16 that

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(\mathbf{d}_{\mathcal{M}}(1NN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}, t'), \mathbf{x}) \geq r\Big) \leq \Big(1 - P_{X \sim \mu(\mathcal{X})}(X \in \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r))\Big)^n. \tag{17}$$

Combining the smoothness of distance metric $\mathcal{M}$, the counterfactual estimation $\hat{y}^{(t')}(\mathbf{x}) = y\Big(1NN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}, t')\Big)$ and Equation 17, we infer that for some ϵ_r corresponding any $r > 0$ we have:

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})| \geq \epsilon_r\Big) \leq \Big(1 - P_{X \sim \mu(\mathcal{X})}(X \in \mathcal{B}_{\mathcal{M}}(\mathbf{x}, r))\Big)^n. \tag{18}$$

Hence, for any arbitrary ϵ we can always find a $\delta(\epsilon, \mathcal{M}, n)$ such that

$$P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})| \geq \epsilon\Big) \leq \delta(\epsilon, \mathcal{M}, n). \tag{19}$$

Note that from Equation 17, we can observe that $\lim_{n\to\infty} 1NN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}, t) \to \mathbf{x}$; this implies *asymptotic convergence of nearest neighbor*. We can also do a similar analysis for KNN for any finite and fixed $K > 0$, however for the sake of simplicity we have shown the finite sample bounds for $1NN$. In contract to previous works on nearest neighbor methods, the result shown Lemma 2 holds for any smooth distance metric, not just for a predefined distance metric.

Lemma 3 *If we can estimate the counterfactual using a finite sample $\mathcal{S}_n \overset{i.i.d}{\sim} \mu(\mathcal{Z}^n)$ such that the true counterfactual $y^{(t)}$ and the estimated counterfactual $\hat{y}^{(t)}(\cdot)$ are farther than ϵ' with probability less than $\delta'(\epsilon', \cdot, n)$ for any given $z \in \mathcal{Z}$ and $t \in \mathcal{T}$, then the estimated individualized treatment $\hat{\tau}(\cdot)$ using a finite sample $\mathcal{S}_n \overset{i.i.d}{\sim} \mu(\mathcal{Z}^n)$ and the true individualized treatment effect $\tau(\cdot)$ are farther than ϵ with probability less than $\delta'(\frac{\epsilon}{2}, \cdot, n)$.*

$$\forall t \in \mathcal{T},\ P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{y}^{(t)}(\mathbf{x}) - y^{(t)}(\mathbf{x})| \geq \epsilon'\Big) \leq \delta'(\epsilon', \cdot, n) \implies P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}\Big(|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \geq \epsilon\Big) \leq \delta'\Big(\frac{\epsilon}{2}, \cdot, n\Big).$$

Proof (Lemma 3). Given that for any $\epsilon' >$, we can find an $\delta^{'}(\epsilon', \cdot, n)$ such that

$$\forall z \in \mathcal{Z},\ \forall t' \in \mathcal{T},\ P_{\mathcal{S}_n \sim \mu(\mathcal{Z}^n)}(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})| \geq \epsilon^{'}) \leq \delta^{'}_{\epsilon'}(\epsilon', \cdot, n).$$

We can further deduce that

$$P\left(\bigvee_{t' \in \mathcal{T}} \Big(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})| \geq \epsilon^{'}\Big)\right) \leq |\mathcal{T}|\ \delta^{'}(\epsilon', \cdot, n). \tag{20}$$

By the triangle inequality, we also know that

$$\sum_{t' \in \mathcal{T}} \Big|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})\Big| \geq \left|\sum_{t' \in \mathcal{T}} \Big(\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})\Big)\right|. \tag{21}$$

Deducting from Equation 20, we have

$$P\left(\sum_{t' \in \mathcal{T}} \Big(|\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})|\Big) \geq |\mathcal{T}|\epsilon^{'}\right) \leq |\mathcal{T}|\ \delta^{'}(\epsilon', \cdot, n).$$

Applying the triangle inequality from Equation 21,

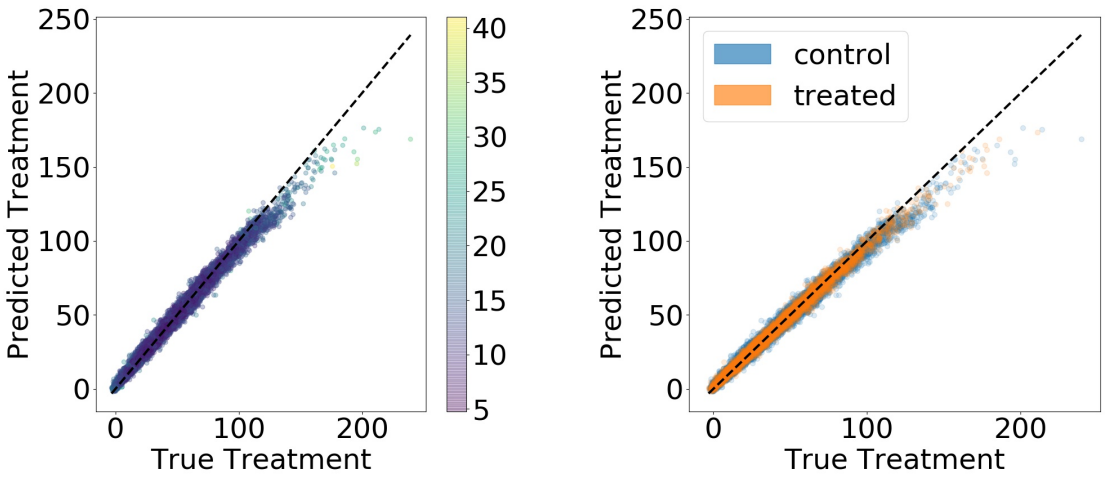
$$P\left(\Big|\sum_{t' \in \mathcal{T}} \Big(\hat{y}^{(t')}(\mathbf{x}) - y^{(t')}(\mathbf{x})\Big)\Big| \geq |\mathcal{T}|\epsilon^{'}\right) \leq |\mathcal{T}|\ \delta^{'}(\epsilon', \cdot, n).$$

Considering the case where $\mathcal{T} = \{0, 1\}$

$$P\left(\Big|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})\Big| \geq 2\epsilon^{'}\right) \leq 2\delta^{'}(\epsilon', \cdot, n).$$

Hence, we can conclude that

$$P\left(\Big|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})\Big| \geq \epsilon\right) \leq 2\delta^{'}\Big(\frac{\epsilon}{2}, \cdot, n\Big).$$



(a) (b)

Figure 9: *Analysis of MALTS performance.* (a) Scatter plot of true treatment effect versus the predicted treatment effect for each unit with the color gradient representing the value of diameter for the corresponding matched group. (b) Scatter plot of true treatment effect versus the predicted treatment effect for each unit with the blue color representing a control unit and the orange color representing a treated unit.

Appendix B

In this section, we further discuss the performance of MALTS in estimating CATEs. We analyze and study the error rate for each estimation unit's matched group with respect to the diameter of the matched group and its treatment assignment.