Variable Importance Matching for Causal Inference

Abstract

Our goal is to produce methods for observational causal inference that are auditable, easy to troubleshoot, yield accurate treatment effect estimates, and scalable to high-dimensional data. We describe a general framework called Model-to-Match that achieves these goals by (i) learning a distance metric via outcome modeling, (ii) creating matched groups using the distance metric, and (iii) using the matched groups to estimate treatment effects. Model-to-Match uses variable importance measurements to construct a distance metric, making it a flexible framework that can be adapted to various applications. Concentrating on the scalability of the problem in the number of potential confounders, we operationalize the Model-to-Match framework with LASSO. We derive performance guarantees for settings where LASSO outcome modeling consistently identifies all confounders (importantly *without* requiring the linear model to be correctly specified). We also provide experimental results demonstrating the auditability of matches, as well as extensions to more general nonparametric outcome modeling.

1 INTRODUCTION

Matching methods are a popular approach to causal inference on observational data due to their conceptual simplicity. These methods aim to emulate randomized controlled trials by pairing similar treated and control units, thus allowing for treatment effect estimation [Stuart, 2010]. One significant benefit of using matching methods is their *auditability* along with their accuracy. Auditability allows domain experts to validate the estimation procedure, argue about the violation of key assumptions, and determine whether the analysis is trustworthy. Parikh et al. [2022a] and Yu et al. [2021] showed that the audit of matched groups using external unstructured data is crucial in high-stakes healthcare and social science scenarios. Since causal analyses often depend on untestable assumptions, it is critical to determine whether all important confounders are accounted for, if data are processed correctly, and whether the treatment and control units in the matched groups are cohesive enough to be comparable [Parikh et al., 2022c]. In high-stakes scenarios, audibility enables domain experts to make data-driven and *trustworthy* decisions and policies.

We would ideally be able to match units that are exactly identical to one another except for treatment assignments [Rosenbaum and Rubin, 1983]. However, exact matches are almost impossible in high-dimensional settings with continuous covariates [Parikh et al., 2022b]. In such scenarios, we aim to create *almost-exact matches* on important covariates. The question then becomes how to construct a distance metric between units that determines who should be in a unit's matched group. We want to learn a distance metric that provides *accurate* causal estimates, ensures *auditability* so we can evaluate and troubleshoot, and is *scalable* to large observational datasets that might be used for high-stakes policy decisions.

We introduce the *Model-to-Match* framework which uses variable importance from prognostic score models to learn a distance metric. The framework has three steps. First, we use machine learning to estimate outcomes and use the measured variable importance to construct a distance metric. Second, we use the learned distance metric to match treatment and control units into matched groups. Third, we use the matched groups to estimate conditional average treatment effects (CATEs). Our research focuses on the first step in this framework, as learning a good distance metric is an essential but difficult step to ensure that matching yields accurate treatment estimates.

A special case of our framework is called *LCM – LASSO Coefficient Matching*. LCM has the characteristics we desire. It is able to accurately estimate treatment effects, creates

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

almost-exact matched groups, and scales better than other comparable methods by orders of magnitude. LCM uses LASSO (Least Absolute Shrinkage and Selection Operator) coefficients to identify important variables and then uses K-nearest neighbors to construct matched groups. LCM benefits from both the efficiency of parametric models and the power of nonlinear modeling by leveraging a parametric method to learn which features to match on and then a nonparametric approach for treatment effect estimation. It is simple to implement yet works extremely well. We perform extensive empirical studies to compare LCM's performance with existing methods. Our results demonstrate that LCM can accurately and efficiently recover true treatment effects even in high-dimensional and non-linear setups without compromising auditability (Section 6). We further propose adaptations of our framework such as (a) metalearner LCM, (b) feature importance matching using decision trees, and (c) LCM-augmented-prognostic scores that work well in complex scenarios (Section 7).

2 BACKGROUND AND ASSUMPTIONS

We study the setting where every individual i in the population $\mathcal{S}$ is assigned to one of the two treatments $T_i \in \{0, 1\}$. Under the stable unit treatment value assumption (SUTVA), we define the potential outcomes of individual i as $Y_i(0)$ and $Y_i(1)$. We consider an i.i.d sample of n individuals, $\mathcal{S}_n$, where we observe (a) p-dimensional pre-treatment covariate vector $\mathbf{X}_i$, (b) assigned treatment T_i, and (c) observed outcome $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$.

The individualized treatment effect is defined as $\tau_i := Y_i(1) - Y_i(0)$. Since we observe only one of the potential outcomes for each unit, τ_i is not observed for any of the units. We need to impute the missing potential outcomes to estimate the treatment effects of interest [Rubin, 2011]. In our setup, we are interested in identifying (a) conditional average treatment effects (CATEs) $\tau(\mathbf{x}) := \mathbb{E}[\tau_i \mid \mathbf{X}_i = \mathbf{x}]$ for all $\mathbf{x} \in Dom(\mathbf{X})$, and (b) the average treatment effect (ATE) $\tau := \mathbb{E}[\tau_i]$.

In observational data (where the treatments are not randomized), the treatment choice and potential outcomes can depend on common variables, which are referred to as confounders. In our setup, we assume that the set of confounders is a subset of the set of pre-treatment covariates, and potential outcomes and treatment assignment are conditionally independent given $\mathbf{X}$: $(Y_i(1), Y_i(0)) \perp T_i \mid \mathbf{X}_i$. This is referred to as conditional ignorability [Rubin, 1974]. Lastly, we assume that the probability of a unit receiving treatment t is bounded away from 1 and 0: $0 < P(T_i = t \mid \mathbf{X}_i = \mathbf{x}) < 1$ – referred to as the positivity assumption. Combining the positivity and conditional ignorability assumptions, adjusting for pre-treatment covariates ($\mathbf{X}$) to identify CATEs and ATE is sufficient.

Matching Methods. Matching methods use a distance-metric, $d_{\mathcal{M}}$, on $\mathbf{X}$ to group similar units with different treatment assignments in order to estimate the causal effects of treatment T on outcome Y. The most popular matching techniques are *propensity score matching* (PSM) [Rosenbaum and Rubin, 1983] and *prognostic score matching* (PGM) [Hansen, 2008]. These techniques project the data to lower dimensions (either a propensity or prognostic score), which are then used for matching. These projections can be sensitive to modeling choices that affect the accuracy of the treatment effect estimates [Kreif et al., 2016]. Further, the units within a matched group can be far from each other in covariate space – i.e., the matched groups are generally not auditable [Parikh et al., 2022b]. To date, the only observational causal inference techniques that attempt to optimize accuracy while maintaining auditability are those stemming from the almost-matching-exactly (AME) framework, namely optimal matching (optMatch) [Yu et al., 2021, Kallus, 2017], genetic matching (GenMatch) [Diamond and Sekhon, 2013], FLAME/DAME [Wang et al., 2017, Dieng et al., 2019], MALTS [Parikh et al., 2022b, 2019, 2022a] and AHB [Morucci et al., 2020] algorithms. FLAME/DAME can scale to extremely large datasets but handles only categorical variables. GenMatch, MALTS, and AHB can also handle both continuous and categorical variables but do not scale as well, thereby limiting their usefulness (see Figure 4 in Section 6). What we develop is a method that yields accurate treatment effect estimates and is auditable like MALTS but can scale to much larger datasets and run at a fraction of the time.

Formally, for a unit i, the K-nearest neighbors of units with treatment t' and corresponding matched group $\text{MG}_{d_{\mathcal{M}}}$ are defined as

$$\text{KNN}_{d_{\mathcal{M}}}(\mathbf{X}_i, t') := \left\{ k : \sum_{j \in \mathcal{S}_n^{(t')}} \mathbb{1} \left[\begin{array}{l} d_{\mathcal{M}}(\mathbf{X}_i, \mathbf{X}_j) \\ < d_{\mathcal{M}}(\mathbf{X}_i, \mathbf{X}_k) \end{array} \right] < K \right\},$$

$$\text{MG}_{d_{\mathcal{M}}}(\mathbf{X}_i) := \bigcup_{t' \in \{0,1\}} \text{KNN}_{d_{\mathcal{M}}}\left(\mathbf{X}_i, t'\right),$$

where $\mathcal{S}_n^{(t')} := \{j : T_j = t'\}$ represents the set of units whose treatment assignment is t'. Match groups can then be used to estimate potential outcomes, $\widehat{Y}_i(t') = \psi\left(\text{KNN}_{d_{\mathcal{M}}}(\mathbf{X}_i, t')\right)$, where ψ is a function of the outcomes of the K-nearest neighbors (e.g. arithmetic mean). The key to accurate estimates (as we will see) is the choice of distance metric. A good choice can lead to interpretable matched groups and accurate treatment effect estimates; a poor choice leads to neither.

3 MODEL-TO-MATCH FRAMEWORK

We propose a framework, called *Model-to-Match*, that focuses on combining prognostic score modeling with distance metric learning for almost-exact matching. Our framework

is divided into three steps: (i) learning the weight matrix $\mathcal{M}$ of a distance metric $d_{\mathcal{M}}$ using a machine learning model, (ii) creating matched groups using the learned $d_{\mathcal{M}}$, and (iii) estimating treatment effects using the matched groups.

We choose our distance metric $d_{\mathcal{M}}$ to be: $d_{\mathcal{M}}(\mathbf{X}_1, \mathbf{X}_2) = d_{\mathcal{M}_c}(\mathbf{X}_{1,c}, \mathbf{X}_{2,c}) + d_{\mathcal{M}_d}(\mathbf{X}_{1,d}, \mathbf{X}_{2,d})$ where $d_{\mathcal{M}_c}(\mathbf{X}_{1,c}, \mathbf{X}_{2,c}) = \|\mathcal{M}_c X_{1,c} - \mathcal{M}_c \mathbf{X}_{2,c}\|_m$; $d_{\mathcal{M}_d}(\mathbf{X}_{1,d}, \mathbf{X}_{2,d}) = \sum_j \mathcal{M}_{d,j,j} \mathbb{1}[\mathbf{X}_{1,d,j} \neq \mathbf{X}_{2,d,j}]$; and $\mathcal{M} = \{\mathcal{M}_c, \mathcal{M}_d\}$. Here, $\mathbf{X}_c$ and $\mathbf{X}_d$ refer to the continuous and discrete components of the covariate vector $\mathbf{X}$.

To learn a distance metric in our Model-to-Match framework we first train two machine learning models, $f^{(0)}$ and $f^{(1)}$ such that for any $i \in \mathcal{S}_n$, $\widehat{Y}_i(t') = f^{(t')}(\mathbf{X}_i)$. We then calculate θ_j, the importance of each covariate $X_{\cdot,j}$ to $f^{(0)}$ and $f^{(1)}$.

Variable Importance Example 1: If the f's are linear estimators, such as LASSO or Ridge where $f^{(t)}_{\boldsymbol{\beta}^{(t)}} = \mathbf{X}\boldsymbol{\beta}^{(t)}$, then θ_j can be $\max_{t'} \boldsymbol{\beta}_j^{(t')}$.

Variable Importance Example 2: If the f's are decision trees then θ_j can be measured via Gini importance, feature permutation importance, or a similar feature importance metric.

Variable Importance Example 3: For f's from backward elimination with ordinary least squares, θ_j can be equal to the drop in R^2 when the j-th feature $X_{\cdot,j}$ is removed.

Variable Importance Example 4: For any generic model class, θ_j can be measured via subtractive model reliance, which measures the change in the loss of a model when a covariate is perturbed [Fisher et al., 2019].

We then set all the diagonal entries, $\mathcal{M}_{j,j}$, in the distance metric $\mathcal{M}$ to be equal to $|\theta_j|$ and all the non-diagonal entries in $\mathcal{M}$ are set to zero. By constructing $\mathcal{M}$ in this way we can interpret each weight, $\mathcal{M}_{j,j}$, as the relative feature importance of covariate $\mathbf{X}_{\cdot,j}$.

We are interested in having an $\mathcal{M}$ that is sparse so that we only match on the important covariates. Further, we want the estimation of f's to be scalable in both the number of samples and the number of covariates. Keeping these requirements in mind, we use ℓ_1-regularized regression, i.e., LASSO, as the modeling method of choice for the majority of the paper. However, our framework is general and can be applied to any supervised model class. For example, we discuss using shallow regression trees to model the f's in Section 7. In practice, LASSO performs well for this step of the framework.

4 LINEAR COEFFICIENT MATCHING

In this section, we operationalize the *Model-to-Match* framework using LASSO [Tibshirani, 1996] as the machine learning algorithm for learning the distance metric and refer to this as LASSO Coefficient Matching (LCM). As in the example in Section 3, we use scaled absolute values of LASSO's coefficients as the diagonal entries for an $\mathcal{M}^*$. Since LASSO's coefficients are sparse, the entries of $\mathcal{M}^*$ will be sparse. This creates a distance metric $d_{\mathcal{M}^*}$ that prioritizes tighter matches on a small number of important covariates, leading to faster runtimes and facilitating matched groups that are close in important covariates.

We perform *honest* causal estimation for a given observed dataset $\mathcal{S}_n$. We divide our samples into two subsets $\mathcal{S}_{n,tr}$ and $\mathcal{S}_{n,est}$. In Step (i), we use $\mathcal{S}_{n,tr}$ to estimate $\boldsymbol{\beta}$'s and, by consequence, learn $d_{\mathcal{M}^*}$. Algorithm 1 describes our training step to learn $\mathcal{M}^*$ using LASSO. In Step (ii), we then perform matching with replacement using $d_{\mathcal{M}^*}$ to get matched groups, $\text{MG}_{d_{\mathcal{M}^*}}(\mathbf{X}_i)$, for each unit $i \in \mathcal{S}_{n,est}$. In Step (iii), we use $\text{MG}_{d_{\mathcal{M}^*}}(\mathbf{X}_i)$ to estimate the CATE for $\mathbf{X} = \mathbf{X}_i$ as $\widehat{\tau}(\mathbf{X}_i) = \widehat{Y}_i(1) - \widehat{Y}_i(0)$ where

$$\widehat{Y}_i(t') = \frac{\sum_{k \in \text{MG}_{d_{\mathcal{M}^*}}(\mathbf{X}_i)} \mathbb{1}[T_k = t'] Y_k}{\sum_{k \in \text{MG}_{d_{\mathcal{M}^*}}(\mathbf{X}_i)} \mathbb{1}[T_k = t']}.$$

Data: Dataset $\mathcal{S}_{n,tr}$
Result: Distance metric $\mathcal{M}^*$
begin
  $W \leftarrow [0, ..., 0] \in \mathbb{R}^p$
  (Loop over treatment possibilities.)
  for t' *in* $\{0, 1\}$ **do**
    (Find units that have treatment t'.)
    $\mathcal{S}^{(t')}_{n,tr} \leftarrow \{i \in \mathcal{S}_{n,tr} : T_i = t'\}$
    (Run LASSO to get coefficients.)
    $\hat{\boldsymbol{\beta}}^{(t')} \leftarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \lambda \|\boldsymbol{\beta}\|_1 + \sum_{i \in \mathcal{S}^{(t')}_{n,tr}} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2$
    (Average the element wise absolute values of the coefficients across treatment and control.)
    for l *in* $\{1, ..., p\}$ **do**
      $W_l \leftarrow W_l + \frac{|\hat{\beta}_l^{(t')}|}{\|\hat{\boldsymbol{\beta}}^{(t')}\|_1}$
    end
  end
  (Coefficients used as stretches in distance metric.)
  $\mathcal{M}^* \leftarrow \mathbf{0}_{p \times p}$
  for l *in* $\{1, ..., p\}$ **do**
    $\mathcal{M}^*_{l,l} \leftarrow \frac{p}{2} W_l$
  end
end

Algorithm 1: Algorithm to estimate $\mathcal{M}^*$ using LASSO

Similar to Chernozhukov et al. [2018], our method performs η-fold cross-fitting by swapping the training set each time to enable the estimation of CATEs for all $i \in \mathcal{S}_n$. Because LASSO does not need many observations to fit the data well, we use only one of the η splits as the training set and the data in the remaining $(1 - \eta)$ splits as the estimation set. Using a smaller amount of data in the learning step allows us to create match groups with a larger portion of the data.

Because the nearest neighbor-based estimation in Step (iii) is local and non-parametric, more data will improve the quality of matched groups and the accuracy of the CATEs.

We employ the Manhattan distance to align with the additive linear form and $\|\cdot\|_1$ regularization of LASSO. In particular, for all $i, j \in \mathcal{S}_{n,est}$ such that $T_i = T_j$, $d_{\mathcal{M}^*} = \sum_{l}^{p} \mathcal{M}^*_{l,l} |\mathbf{X}_{i,l} - \mathbf{X}_{j,l}|$. Our method has three hyperparameters: η, λ, and K. We learn λ using cross-validation in our training in Step (i). The number of nearest neighbors, K, and the number of splits for cross-fitting, η, can be chosen through cross-validation or set manually.

5 THEORETICAL RESULTS

Here, we prove optimality properties of using LASSO to learn our distance metric. We then show under what conditions LCM guarantees consistency in CATE estimation. Proofs are included in the Supplementary Materials.

Theorem 5.1 motivates LCM. It shows that if the potential outcomes are linear in the predictors then using the absolute values of the coefficients in these models as the stretches in a distance metric guarantees that as the distance between two samples decreases, their expected outcomes become closer.

Theorem 5.1. *[Closeness in $\mathbf{X}$ implies closeness in Y]. Consider a p-dimensional covariate space where for $t' \in \{0,1\}$, $f^{(t')}(\mathbf{X}_i) = \mathbb{E}[Y_i|\mathbf{X} = \mathbf{X}_i, T = t'] = \mathbf{X}_i \boldsymbol{\beta}^{(t')}$. Construct $\mathcal{M} \in \mathbb{R}^{p\times p}$ where for all $l, r \in \{1, ..., p\}$ $\mathcal{M}_{l,l} = |\beta_l^{(t')}|$ and for $l \neq r$ $\mathcal{M}_{l,r} = 0$. Then, $\forall i, j$, we have that $d_{\mathcal{M}}(\mathbf{X}_i, \mathbf{X}_j) \geq \left| f^{(t')}(\mathbf{X}_i) - f^{(t')}(\mathbf{X}_j) \right|$.*

From here, we define a diagonal Mahalanobis distance matrix as any $\widetilde{\mathcal{M}} \in \mathbb{R}^{p\times p}$ that is diagonal (for all $l, r \in \{1, ..., p\}$, $l \neq r$, $\widetilde{\mathcal{M}}_{l,r} = 0$) and has non-negative entries ($\widetilde{\mathcal{M}}_{l,l} \geq 0$). We show in Theorem 5.2 that the $\mathcal{M}$ from Theorem 5.1 is the optimal stretch matrix, compared to any other equally scaled diagonal Mahlanobis distance matrix, in regards to the maximum absolute difference in expected outcomes.

Theorem 5.2. *[Optimality of $\mathcal{M}$] Using the setup of Theorem 5.1, let $\text{supp}(\mathbf{X}) = \mathbb{R}^p$. Consider an arbitrary diagonal Mahalanobis distance matrix $\widetilde{\mathcal{M}} \in \mathbb{R}^{p\times p}$ where $\|\widetilde{\mathcal{M}}\|_1 = \|\boldsymbol{\beta}^{(t')}\|_1$ and $\widetilde{\mathcal{M}}_{l,l} > 0$ when $|\beta_l^{(t')}| > 0$. For some $\epsilon \geq 0$ and $\mathbf{X}_1 \in \mathbb{R}^p$, define $S_{\widetilde{\mathcal{M}},\epsilon}(\mathbf{X}_1) := \{\mathbf{X}_2 : \mathbf{X}_2 \in \mathbb{R}^p, d_{\widetilde{\mathcal{M}}}(\mathbf{X}_1, \mathbf{X}_2) = \epsilon\}$. Then,*

$$\sup_{\mathbf{X}_2 \in S_{\mathcal{M},\epsilon}(\mathbf{X}_1)} |f^{(t')}(\mathbf{X}_1) - f^{(t')}(\mathbf{X}_2)| \leq \sup_{\mathbf{X}_3 \in S_{\widetilde{\mathcal{M}},\epsilon}(\mathbf{X}_1)} |f^{(t')}(\mathbf{X}_1) - f^{(t')}(\mathbf{X}_3)|.$$

These results show how a linear outcome model induces a meaningful distance metric for causal inference. The following theorem states that when we do not know the true value of the coefficients (and more generally when the model is non-linear but LASSO still recovers its support), we can employ the LCM procedure of Section 4 to generate a distance metric that yields consistent estimates of CATEs. This theorem uses variable importance, discussed in Section 3.

Theorem 5.3. *[Consistency of LCM] For $t' \in \{0,1\}$, let $f^{(t')}(\mathbf{X}_i) = \mathbb{E}[Y_i|\mathbf{X} = \mathbf{X}_i, T = t']$. Let $f^{(t')}$ be Lipschitz continuous and,*

$$\text{supp}\left(f^{(t')}\right) := \left\{ j : \mathbf{X}_{\cdot,j} \text{ is important in } f^{(t')} \right\}.$$

Denote $d_{\mathcal{M}^*}$ as the distance metric learned by LCM in Section 4 and let $\text{supp}(\mathcal{M}^*) = \{j : \mathcal{M}^*_{j,j} > 0\}$. LCM is consistent for CATE estimation if $\text{supp}\left(f^{(0)}\right) \bigcup \text{supp}\left(f^{(1)}\right) \subseteq \text{supp}(\mathcal{M}^*)$.

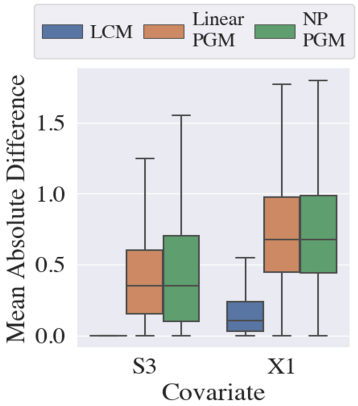
This result follows from LASSO and its adaptations' ability to estimate sparse coefficient vectors in high dimensions, even when $n < p$ [Meinshausen and Yu, 2009, Zhou, 2010, Wasserman and Roeder, 2009, Meinshausen and Bühlmann, 2006]. LASSO also exhibits consistency for feature selection in some nonlinear settings [Zhang et al., 2016].

6 EXPERIMENTAL RESULTS

Our experiments focus on factors crucial in high-stakes causal inference. **(i) Accuracy and Auditability**: We compare LCM's matched groups to PGM's and highlight the importance of auditability. **(ii) Nonlinear Outcome**: We study if LCM is sensitive to model misspecification and compare our results to linear PGM (which uses the same underlying prognostic model as LCM). **(iii) Scalability**: We compare LCM to existing AME algorithms in both runtime and estimation performance as both the number of observations and the number of features increase.

6.1 ACCURACY AND AUDITABILITY

Matching enables us to investigate whether a CATE is estimated in a trustworthy manner by *auditing* the quality of the matched groups. We now highlight how LCM produces accurate estimates while matching tightly on important covariates. We work with the ACIC 2018 Atlantic Causal Inference Conference semi-synthetic dataset [Carvalho et al., 2019], which is based on data from the National Study of Learning Mindsets randomized trial [Yeager, 2015-2016]. The dataset contains 10,000 students across 76 schools. There are four categorical student-level covariates and one categorical and five continuous school-level covariates. Carvalho et al. [2019] constructed this semi-synthetic dataset by



(a)

	T	S3	X1
Query	1	5	-.05
LCM	0	5	-.05
	0	5	-.05
	0	5	-.05
	0	5	-.05
Linear PGM	0	5	-.05
	0	5	-.05
	0	**6**	**1.26**
	0	**6**	**1.26**
NP PGM	0	5	-.05
	0	5	-.05
	0	**6**	**.98**
	0	**6**	**.98**

(b)

Figure 1: Closeness in important covariates for matched groups produced by LCM, linear PGM, and nonparametric (NP) PGM. *(a)* shows the mean absolute difference between a query unit and its matched group's covariate values. Smaller values imply better and tighter matches. *(b)* shows, for a random sample, the four nearest neighbors of opposite treatment under LCM, linear PGM, and NP PGM. In *(b)*, the text in red indicates values that are far from the query's value. S3 indicates the self-reported prior achievements of students and is important for selection into treatment, and X1 indicates school-level average mindset score of the students and is an effect modifier.

drawing covariates from the real experiment and then synthetically generating treatment assignments and outcomes. Details can be found in Carvalho et al. [2019].

We ran our method alongside linear PGM, computed using LASSO, and nonparametric PGM, computed using gradient boosted trees. All three methods recovered ATE estimates that were close to the true value of 0.24 – **LCM**: 0.249, **Linear PGM**: 0.251, and **Nonparametric PGM**: 0.260, which is also in line with the estimates of other interpretable and uninterpretable methods described in Carvalho et al. [2019].

While all three methods accurately estimate the ATE, *only LCM matches almost exactly on important covariates*. We compare how tightly LCM, linear PGM, and nonparametric PGM fit on a covariate that is identified as important for selection into treatment (S3) and one that is an effect modifier (X1). Figure 1 shows that LCM matches tighter on important covariates than PGM. In this way, LCM more closely emulates exact-matching and results in more intuitive and auditable match groups. The fact that LCM accurately estimates the treatment effect and matches so tightly on these important covariates increases the trust we have in our conclusions. We expand on these findings and show that LCM matches tighter across all the effect modifiers in the Supplementary Materials.

6.2 NONLINEAR OUTCOMES

We have shown that linear prognostic score matches are not tight on important covariates, leading to unintuitive matched groups. However, LASSO estimated prognostic scores are more interpretable than scores estimated with gradient boosted trees. This interpretability comes at a cost: the performance of linear PGM heavily depends on the linearity of the underlying data generation process. LCM is more robust to nonlinear data because its LASSO component is used *only to determine the relative weight of features* in the distrance metric (not to model the outcome with a linear combination of the covariates).

We compare CATE estimation accuracy of LCM and linear PGM on two synthetically generated datasets where the outcome is a non-linear function of the covariates. We call these datasets **Sine** and **Exponential** to align with their underlying potential outcome functions. We generate 5000 samples and 100 covariates for each dataset. For **Sine**, the outcome function is

$$Y_i = \sin(X_{i,1}) - T_i \sin(X_{i,2}).$$

Whereas, for **Exponential**, the outcome function is

$$Y_i = 2e^{X_{i,1}} - \sum_{j=2}^{3} e^{X_{i,j}} + T_i e^{X_{i,4}}.$$

We outline the specific details of the data generation processes in the Supplementary Materials. Figure 2 shows that LCM is more robust to nonlinear outcome functions than linear PGM.

Again, the good performance of LCM is unsurprising because it performs nonlinear estimation in Step (iii), using the linear LASSO method in Step (i) only to pinpoint important covariates upon which nonlinear estimation can be successfully performed.

6.3 SCALABILITY

Existing almost-matching-exactly methods learn covariate weights and/or create match groups through computationally expensive and data hungry optimization algorithms. In this section we compare LCM to MALTS [Parikh et al., 2022b], GenMatch [Diamond and Sekhon, 2013], and AHB [Morucci et al., 2020] in regards to scalability in runtime and in CATE estimation accuracy. We omit FLAME/DAME [Wang et al., 2017, Dieng et al., 2019] from this comparison since it can only handle discrete covariates.

We generate synthetic datasets of various sizes from the quadratic data generation process described in Parikh et al. [2022b] and the Supplementary Materials. We first measure the runtime scalability of LCM, MALTS, GenMatch, and AHB with respect to the number of samples, n, and number

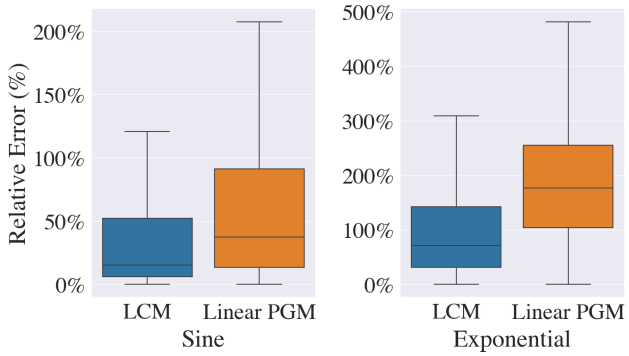


Figure 2: Comparing LCM's and Linear PGM's performances for nonlinear synthetically generated datasets **Sine** and **Exponential**.

of covariates, p. To measure scaling runtime in n, we keep the number of covariates constant at 64 and increase the number of samples from 256 to 8192. To measure scaling in p, we set the number of samples to be 2048 and vary the number of covariates from 8 to 1024. The Supplementary Materials contain further information on how runtimes were measured. Figure 4 shows the runtimes for each of the AME algorithms on these various dataset sizes, highlighting the multiple-order-of-magnitude runtime disparity between LCM and other AME methods. MALTS ran out of memory (16GB RAM) for the largest dataset in each plot. We stopped increasing the dataset sizes for AHB when its runtime surpassed the longest measured runtime of all other methods.

As discussed in Section 4, LASSO is capable of recovering sparse βs and important features in high dimensional settings. Naturally, LCM also excels at producing accurate CATE estimates as the number of irrelevant covariates grows. Figure 3 shows how LCM is robust to added noise as the number of unimportant covariates grows – unlike MALTS and GenMatch, which struggle to learn an accurate distance metric as the dimensionality of the covariate space increases. Here, we keep the number of important covariates equal to 8.

7 MODEL-TO-MATCH ADAPTATIONS

In this section, we propose three adaptations of the Model-to-Match framework that extend LCM. The first approach uses a metalearner variant of LCM and shows improvement in CATE estimation in certain settings. The second adaptation proposes the use of a tree-based outcome modeling approach in place of LASSO. The third adaptation combines prognostic score matching with LCM to yield accurate CATEs and tight match groups.

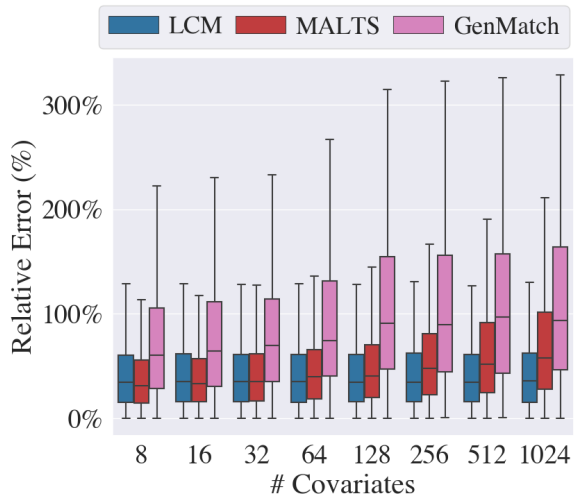


Figure 3: Absolute CATE estimation error relative to the true ATE for LCM, MALTS, and GenMatch as the number of covariates increases.

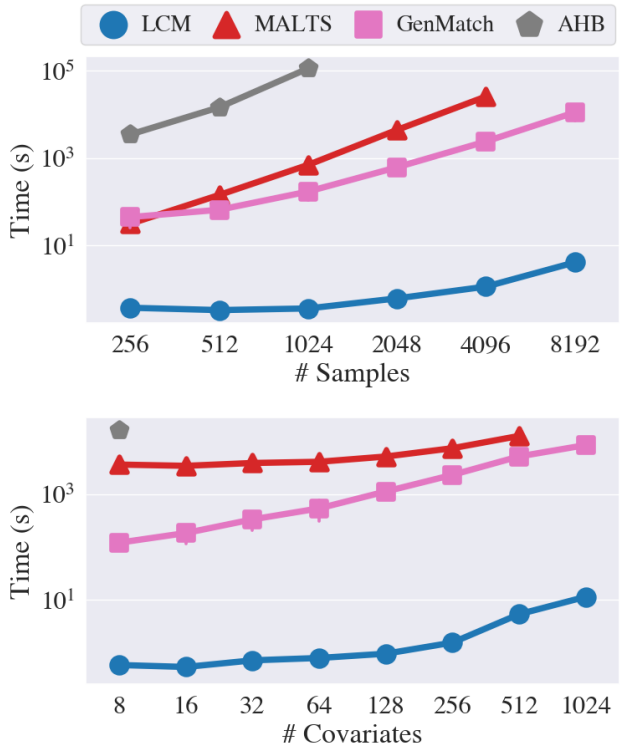


Figure 4: Scalability in n and p for LCM, MALTS, GenMatch, and AHB. Error bars represent the standard deviation of runtimes across 20 iterations.

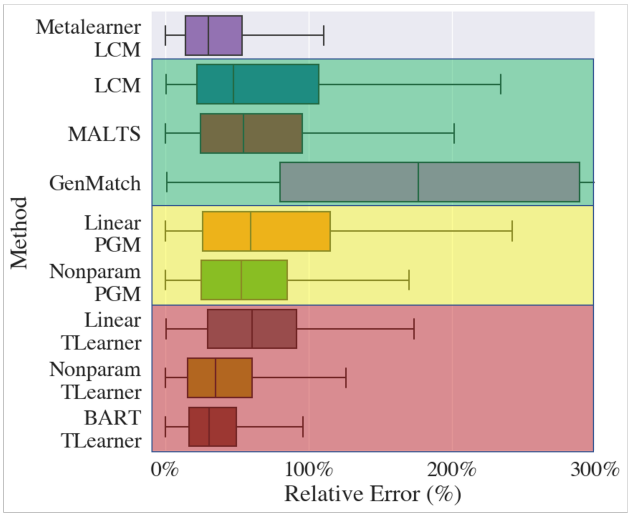


Figure 5: Absolute CATE estimation error relative to the true ATE for various methods on the **Sine** data generation process. The transparent boxes separate the methods into different categories. **Green**: Almost exact matching methods. **Yellow**: Other matching methods. **Red**: TLearner methods.

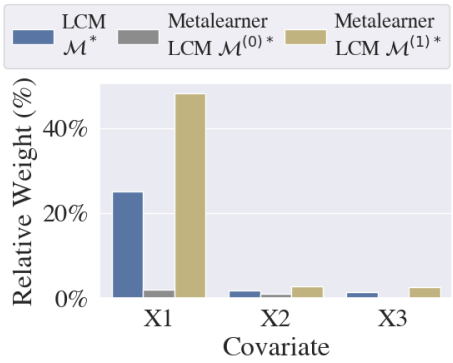


Figure 6: Relative covariate weights averaged over the η-folds for LCM $\mathcal{M}^*$, Metalearner LCM $\mathcal{M}^{(0)*}$, and Metalearner LCM $\mathcal{M}^{(0)*}$. This shows that the Metalearner LCM's distance metrics are different between treatment and control groups.

7.1 METALEARNER LCM

Metalearners leverage powerful regression tools for estimating heterogenous treatment effects [Künzel et al., 2019]. LASSO Coefficient Matching can be adapted to run similar to the T-learner outlined by Künzel et al. [2019] by learning separate distance metrics for control and treated units. The metalearner adaptation of LCM is advantageous when certain covariates have vastly different effects on the outcome depending on if a sample received treatment or not.

For Metalearner LCM, we learn a separate distance metric, $d_{\mathcal{M}^{(t')*}}$, for each $t' \in \{0, 1\}$. Specifically, for $l, r \in \{1, \ldots, p\}$ we set $\mathcal{M}_{l,l}^{(t')*} = |\hat{\beta}_l^{(t')}| \frac{p}{\|\hat{\boldsymbol{\beta}}^{(t')}\|_1}$, where $\hat{\boldsymbol{\beta}}^{(t')} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \lambda \|\boldsymbol{\beta}\|_1 + \sum_{i \in \mathcal{S}_{n,tr}^{(t')}} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2$, and $\mathcal{M}_{l,r}^{(t')^*} = 0$ when $l \neq r$. In Step (ii), for each unit $i \in \mathcal{S}_{n,est}$, we find K-nearest neighbors with replacement using the corresponding distance metric in each treatment arm.

To illustrate the advantage of the Metalearner LCM, we consider the same **Sine** data generation process used in Section 6.2. In **Sine**, covariate $X_{i,1}$ is important to the outcome under both treatment regimes ($Y_i(0)$ and $Y_i(1)$) while covariate $X_{i,2}$ is only relevant to the outcome under treatment ($Y_i(1)$). We generate 500 samples and 10 covariates. We compare LCM to the previously used matching methods along with linear and nonparametric T-Learners. Figure 5 shows estimated CATE errors for each method. Metalearner LCM improves upon LCM, which already outperforms other matching methods, and is comparable to T-Learners.

Figure 6 shows how Metalearner LCM stretches the control and treatment response surfaces differently, whereas regular LCM learns a global metric that is a linear combination of the two treatment spaces. The Metalearner variant is more suitable for problems in which accurate CATE estimation is more important than emulating a randomized experiment.

7.2 FEATURE IMPORTANCE MATCHING

LASSO can often find the important features, even if the true data generation process is nonlinear [Zhang et al., 2016]. However, in cases where it cannot, we can use any nonlinear method (decision tree, random forest, BART, AdaBoost, etc.) from which we can extract a measure of feature importance. These feature importance values can be used in place of LASSO coefficients in Algorithm 1 as weights for matching.

We demonstrate using shallow decisions trees as the model and Gini importance as the feature importance measure [Menze et al., 2009]. We use a shallow decision tree to promote sparsity and account for nonlinearities in the outcome space. We generate a dataset with 1000 samples and 10 covariates where only the first covariate is important: $Y_i(0) = X_{i,1}^2 + \epsilon_{i,y}$ and $Y_i(1) = X_{i,1}^2 + 10 + \epsilon_{i,y}$. A linear approach will not find this important covariate because it is symmetric around 0. The full data generation process is outlined in the Supplementary Materials. Figure 7 shows that in this setting, the tree-based method creates more accurate CATE estimates than the LASSO method (LCM).

7.3 LCM-AUGMENTED-PGM

As shown in Section 6.1, LCM produces tighter matched groups on important covariates than linear and non-

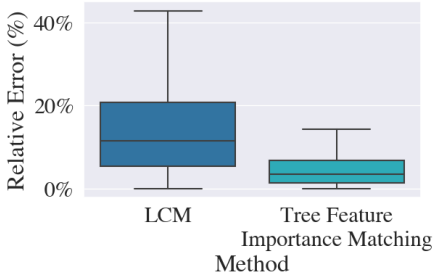


Figure 7: Absolute CATE estimation error relative to the true ATE for LCM vs. classification and regression decision tree (CART) Gini feature importance.

parametric PGM. However, PGM sometimes can estimate CATEs more accurately while not producing tight matched groups. This might occur either when the parametric prognostic model is correctly specified or when there is a strong non-linear effect that a non-parametric prognostic score can model accurately. In such situations, we propose augmenting PGM with LCM to guarantee tight matches and accurate CATE estimates. Our LCM-augmented-PGM (LAP) is a two stage procedure. In the first stage, we match using prognostic scores and create large matched groups. In the second stage, we match using the distance metric learned via LCM inside each PGM matched group. The first stage leverages the flexibility of outcome modeling and the second stage ensures tight matching on important covariates.

We compare LCM and LAP using the quadratic data generation process used in Section 6.3 and described in the Supplementary Materials. We generate 5000 units and 20 continuous covariates, of which the first 5 are important and the other 15 are irrelevant. Here, we first do 25 nearest neighbors matching with PGM and then perform 5 nearest neighbors matching using the LCM learned distance metric. Figure 8 shows that for this problem setup, both linear LAP and non-parametric LAP are more accurate than LCM. Further, Figure 9 shows that the matches created using non-parametric LAP are almost equally as tight as LCM's matches on the 5 important covariates and do not prioritize matching on irrelevant covariates.

8 DISCUSSION AND CONCLUSION

Model-to-Match is a fast, scalable, and auditable framework for observational causal inference. Unlike other almost-matching-exactly approaches, Model-to-Match can scale to large datasets and high-dimensional settings and is flexible in regards to how the outcome space is modeled to learn a distance metric. We implemented Model-to-Match using LASSO as our machine learning algorithm of choice and refer to this as LCM. We show many desirable properties of LCM – including robustness to model misspecification and ability to handle high-dimensional settings – and provide details on its consistency for CATE estimation.

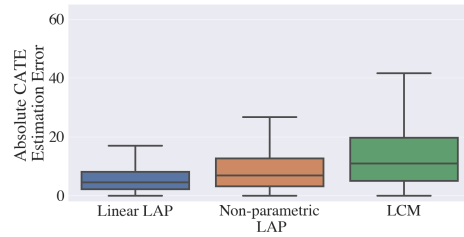


Figure 8: Absolute CATE estimation error for linear LAP (blue), non-parametric LAP (orange) and LCM (green).

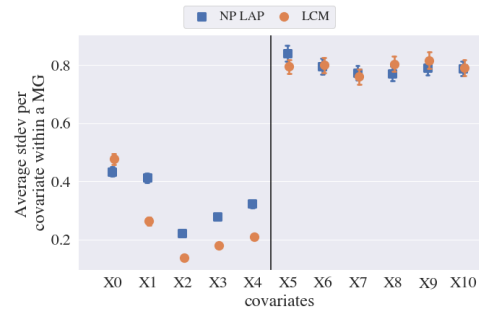


Figure 9: Average standard deviation for each covariate inside the matched groups for NP LAP and LCM. The smaller the standard deviation, the tighter the match on those covariates. The dataset has 20 covariates, but we only show 11 (5 important and 6 unimportant) for ease of presentation.

Limitations and Future Directions. Model-to-Match is for iid data and should be extended to situations with either network interference or time-series effects. Furthermore, Model-to-Match is sensitive to the choice of variable importance metric – leading to confounding bias if the correct support is not recovered. While we introduce our framework for categorical treatments, we are working on extending its application to continuous treatment regimes.

Other variations of Model-to-Match are easily possible. While we show sparse decision trees as a potential substitute to LASSO, any machine learning algorithm can be used. Furthermore, one can use other configurations in the matching and estimation steps of the framework, such as using a $\|\cdot\|_2$ norm instead of $\|\cdot\|_1$, employing a caliper matching method instead of K nearest neighbors, or choosing a different post-matching estimator instead of arithmetic average for potential outcomes. This level of flexibility makes Model-to-Match a framework that can be adapted to a variety of practical problems. In future work, we plan to both study the theory behind different Model-to-Match variations and implement our framework on large, real-world datasets such as electronic health records, genome studies, living standards measurement studies, etc.

References

Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge, 2019. URL https://arxiv.org/abs/1907.07592.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Alexis Diamond and Jasjeet S. Sekhon. 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, 2013.

Awa Dieng, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Interpretable almost-exact matching for causal inference. *Proceedings of Machine Learning Research (Proceedings of AISTATS)*, 89:2445, 2019.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177): 1–81, 2019.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

Nathan Kallus. A Framework for Optimal Matching for Causal Inference. In Aarti Singh and Jerry Zhu, 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, 20–22 Apr 2017.

Noémi Kreif, Susan Gruber, Rosalba Radice, Richard Grieve, and Jasjeet S Sekhon. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research*, 25(5):2315–2336, 2016. doi: 10.1177/0962280214521341.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019. doi: 10.1073/pnas.1804597116. URL https://doi.org/10.1073/pnas.1804597116.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), June 2006. doi: 10.1214/009053606000000281. URL https://doi.org/10.1214/009053606000000281.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1), February 2009. doi: 10.1214/07-aos582. URL https://doi.org/10.1214/07-aos582.

Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10:1–16, 2009.

Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Adaptive hyper-box matching for interpretable individualized treatment effect estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1089–1098. PMLR, 2020.

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

Harsh Parikh, Kentaro Hoffman, Haoqi Sun, Sahar F. Zafar, Wendong Ge, Jin Jing, Lin Liu, Jimeng Sun, Aaron F. Struck, Alexander Volfovksy, Cynthia Rudin, and M. Brandon Westover. Effects of epileptiform activity on discharge outcome in critically ill patients: A retrospective cross-sectional study. 2022a.

Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. MALTS: Matching after learning to stretch. *Journal of Machine Learning Research*, 23(240):1–42, 2022b.

Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17346–17358. PMLR, 17–23 Jul 2022c.

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

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

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2011.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tianyu Wang, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. FLAME: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*, 2017.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A), October 2009. doi: 10.1214/08-aos646. URL https://doi.org/10.1214/08-aos646.

David S. Yeager. The national study of learning mindsets [united states]. 2015-2016. URL https://doi.org/10.3886/ICPSR37353.v4.

Ruoqi Yu, Dylan S. Small, David Harding, José Aveldanes, and Paul R. Rosenbaum. Optimal matching for observational studies that integrate quantitative and qualitative research. *Statistics and Public Policy*, 8(1):42–52, 2021. doi: 10.1080/2330443X.2021.1919260. URL https://doi.org/10.1080/2330443X.2021.1919260.

Yue Zhang, Soumya Ray, and Weihong Guo. On the consistency of feature selection with lasso for non-linear targets. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, page 183–191, 2016.

Shuheng Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation, 2010. URL https://arxiv.org/abs/1002.1583.