Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-022-00886-5

Differentiated matching for individual and average treatment effect estimation

Zhao Ziyu[1] · Kun Kuang[1] · Bo Li[2] · Peng Cui[2] · Runze Wu[3] · Jun Xiao[1] · Fei Wu[4,5,6]

Received: 1 March 2021 / Accepted: 25 October 2022

Abstract

One fundamental problem of causal inference is estimating treatment effect with observational data where variables are confounded. The traditional way of controlling the confounding bias is to match units with different treatments but similar variables. However, traditional matching methods fail on selection and differentiation among the pool of numerous potential confounders, leading to possible under-performance. In this paper, we give a theoretical analysis of confounder differentiation and propose a novel Differentiated Matching (DM) algorithm for both individual and average treatment effect estimation by learning confounder weights for variable differentiation and unit matching. To address the distribution shift in confounder weights learning, we further propose a Propensity Score based DM (PSDM) algorithm by weighted regression with the inverse of the propensity score. Extensive experiments on both synthetic and real-world datasets demonstrate that the proposed algorithms achieve better performance than other matching methods on treatment effect estimation.

Keywords Treatment effect · Matching method · Confounder differentiation · Causal inference

Responsible editor: Sriraam Natarajan

These authors contributed equally to this work

✉ Kun Kuang
kunkuang@zju.edu.cn

✉ Fei Wu
wufei@cs.zju.edu.cn

Extended author information available on the last page of the article

Published online: 09 November 2022

1 Introduction

Causal inference (Holland 1986; Kuang et al. 2020) is a powerful statistical modeling tool for explanatory analysis in many decision-making applications. Treatment effect estimation is a fundamental problem for causal inference. Its main challenge is to remove the confounding bias induced by the different variables' distribution between treated and control units. The golden standard approach for removing confounding bias is randomized experiments (Lewis and Reiley 2008), where different treatments are randomly assigned to units. In real-world applications, however, fully randomized experiments are usually extremely expensive (Kohavi and Longbotham 2011) or sometimes even infeasible (Bottou et al. 2013). Hence, it is paramount to infer treatment effects from observational data with confounding bias.

Matching is a classic approach to treatment effect estimation, especially individual treatment effect estimation in observational studies, which reduces the confounding bias by matching units with different treatments but similar covariates. Suppose an individual taking medication wanted to determine whether the medicine caused recovery (individual treatment effect). The patients taking medication would compare their outcomes with a unit in the control group with similar covariates. Characterizing the similarity between individuals is a central problem for matching methods. For traditional matching methods, the distance metric is given in advance rather than learned (Rosenbaum and Rubin 1983; Iacus et al. 2012; Kallus 2017, 2019). These methods match units based on all observed variables, which may lead to incorrect matches based on irrelevant variables. For example, the effect of taking medication is often not directly related to marital status. However, traditional matching methods also consider these unrelated variables, leading to poor performance in estimating the treatment effect.

Some methods have also considered learning the appropriate distance metric for matching in recent years (Diamond and Sekhon 2013; Liu et al. 2019; Wang et al. 2021). Nevertheless, these methods tend to concentrate only on selecting confounders and ignore their differentiation since different confounders may have an unequal effect on the outcome, resulting in poor performance in real-life scenarios.

In this paper, we focus on both individual and average treatment effect estimation via the matching method, taking into account the differentiation of confounders. Moreover, we consider the distribution shift problem in confounder weights learning and introduce propensity score into confounder weights learning. The main contributions of this paper are as follows:

- We study a problem of individual and average treatment effect estimation via the matching method while considering the confounder differentiation and distribution shift problem of confounder weights learning.
- We propose a novel Differentiated Matching (DM) algorithm to jointly select confounders, optimize confounder weights for weighted matching on units and simultaneously estimate the individual and average treatment effect.
- We further propose a Propensity Score based DM (PSDM) to address the distribution shift problem of confounder weights learning with a theoretical guarantee.

- Experiments on both synthetic and real-world datasets show the superior performance of our proposed algorithms on both individual and average treatment effect estimation compared with previous matching methods.

2 Related work

Matching methods The matching method aims to match individuals with similar variables but different treatment assignments in causal literature. Exact matching aims to find the "identical twins" that obtain different treatments but with exactly identical variables. However, exact matching is not possible in high dimensions or when dealing with continuous variables. Then, coarsen exact matching (Iacus et al. 2012) was proposed by recording variables to binary. Propensity score matching (Rosenbaum and Rubin 1983) was proposed to reduce the dimension of variables by projecting entire variables to one dimension. Thus it cannot be applied to estimate the individual treatment effect. Similarity, Rosenbaum (2017) introduced an "optimal matching" to reduce the dimension via a pre-defined distance measure. Kallus (2017), Kallus (2019) proposed a encompassing framework for optimal matching estimators for causal inference. These methods, however, matched all variables without confounder selection and differentiation, resulting in poor performance in high-dimensional settings.

Recently, some metric-based algorithms were proposed to enhance the ability to estimate the treatment effect. GenMatch algorithm (Diamond and Sekhon 2013) was proposed to learn variables' weights for a weighted matching via evolutionary search, and the algorithm matches by Mahalanobis distance with addition variables weight. However, its complexity grows exponentially with the number of observed variables. Wang et al. (2021) proposed FLAME (Fast Large-scale Almost Matching Exactly) algorithm to improve the coarsened exact matching (Iacus et al. 2012) to learn a suitable distance metric and dropping covariates to permit matches. Liu et al. (2019) further proposed DAME (Dynamic Almost Matching Exactly) to improve the matching performance compared with FLAME (Wang et al. 2021). These methods considered confounder selection and proposed a weighted Hamming distance for matching, but they ignore the differentiation of confounders. Moreover, these methods only focus on categorical variables and cannot be applied to matching with continuous variables.

Confounder differentiation Kuang et al. (2019) introduced confounder differentiation and applied it to balancing methods. However, it can only be applied to estimate the average treatment effect and cannot estimate the individual treatment effect. Moreover, it directly learns the confounders' weights through a simple linear regression from the control group to its outcome. When facing a more complicated scenario, the distribution between the treated and control groups can be highly different. The learned confounders' weights cannot be applied directly to the treated group.

In this paper, we prove the necessity of confounder selection and differentiation for the matching method and propose a differentiated confounder matching algorithm for treatment effect estimation. At the same time, we consider the problem of distribution shift between treated and control groups and propose a propensity score based differentiated matching algorithm to solve the problem.

Table 1 Symbols and definitions

Symbol	Definition
n_t (n_c)	Sample size for treated (control) group
p	Dimension of observed variables
$T \in \mathbb{R}^{n\times 1}$	Treatment
$Y \in \mathbb{R}^{n\times 1}$	Outcome
$\mathbf{X} \in \mathbb{R}^{n\times p}$	Observed variables
$(i, j)\|T_i = 1, T_j = 0$	Matched treated-control units pair
$S = \{(i, j)\|T_i = 1, T_j = 0\}$	Set of matched units pair

3 Problem and assumption

Our goal is to estimate the treatment effect, including individual and average treatment effects from observational data. With the potential outcome framework proposed by Imbens and Rubin (2015), we define the treatment as a random variable T and a potential outcome as $Y(t)$ which corresponds to a specific treatment assignment $T = t$. In this paper, we focus on estimating the causal effect of binary treatment, which is $t \in \{0, 1\}$. We define the units which received treatment ($T = 1$) as treated units and the other units with $T = 0$ as control units. Then, for each unit indexed by $i = 1, 2, \cdots, n$, we observe a treatment T_i, an outcome Y_i^{obs} and a vector of observed variables $X_i \in \mathbb{R}^{p\times 1}$, where the observed outcome Y_i^{obs} of unit i denotes by:

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{1}$$

The numbers of treated and control units are equal to n_t and n_c, and the dimension of all observed variables is p. Table 1 summarized the symbols and definitions.

The essential goal of causal inference in observational studies is to evaluate the causal effect of treatment T on outcome Y, including Individual Treatment effect on Treated (ITT) and Average Treatment effect on Treated (ATT). ITT of a treated unit i refers to the difference between the potential outcome of units i under treated and control status. Formally, the ITT of treated unit i is defined as:

$$ITT_i = Y_i(1) - Y_i(0). \tag{2}$$

ATT represents the average ITT over all treated units. Formally, the ATT is defined as:

$$ATT = E(ITT_i) = E(Y(1) - Y(0)|T = 1), \tag{3}$$

where $Y(1)$ and $Y(0)$ represent the potential outcome of units with treatment status as treated $T = 1$ and control $T = 0$, respectively. $E(\cdot)$ refers to the expectation function.

The Eqs. (2) and (3) are infeasible because of the counterfactual problem (Chan et al. 2010), that for each treated unit i, we can only observe one of the two potential outcomes $Y_i(1)$, the other potential outcome $Y_i(0)$ is unobserved or counterfactual. One can address this counterfactual problem by approximating the unobserved potential outcome. The most straightforward approach is directly comparing the average outcome between the treated and control units. However, in observational studies, directly comparing two samples is likely to have a bias if the treatment assignment is not random, as confounding bias is not taken into account (Chan et al. 2010).

In this paper, we focus on matching methods for treatment effect estimation, including ITT_i and ATT. Specifically, for each treated units i, we try to find a control unit j to match with guarantee that the bias between the match units pair (i, j) is removed. Then, we can estimate the ITT of treated unit i as:

$$\widehat{ITT}_i = Y_i^{obs} - Y_j^{obs}. \tag{4}$$

With the matched treated-control units pair set $S = \{(i, j)|T_i = 1, T_j = 0\}$, one can easily evaluate the ATT as follows:

$$\widehat{ATT} = \sum_{(i,j)\in S} \frac{1}{n_t}(Y_i^{obs} - Y_j^{obs}). \tag{5}$$

Throughout this paper, we assume following standard assumptions (Rosenbaum and Rubin 1983) are satisfied.

Assumption 1 **Stable Unit Treatment Value.** Given the observed variables, the distribution of potential outcomes for one unit is assumed to be unaffected by the particular treatment assignment of another unit.

Assumption 2 **Unconfoundedness.** Given the observed variables, the distribution of treatment is independent of potential outcome. Formally, $T \perp \big(Y(0), Y(1)\big)|\mathbf{X}$.

Assumption 3 **Overlap.** Every unit has a nonzero probability to receive either treatment status when given the observed variables. Formally, $0 < p(T = 1|\mathbf{X}) < 1$.

4 Differentiated matching estimator

In this section, we first introduce the traditional matching methods and analysis their limitations in the scenario of big data. Then, we propose a differentiated matching method with a theoretical analysis.

4.1 Traditional matching method

Matching is a classic approach for treatment effect estimation in observational studies, including exact matching, coarsened exact matching and many distance metric based matching methods. Specifically, for each treated unit i, these methods try to find its closest match among control units as follows:

$$match(i) = \arg\min_{j:T_j=0} d(X_i, X_j), \tag{6}$$

where function $d(a, b)$ measures the distance between the vector of variables a and b. Based on Eq. (6), exact matching constrains the match unit pair (i, j) to satisfy that $X_i = X_j$, and other distance based matching methods focus on proposing more appropriate distance metric $d(\cdot)$ for units matching. However, irrelevant covariates and unimportant covariates may dominate the distance metric for matching, leading to poor performance in high dimensional settings.

4.2 Differentiated matching

To address the problem of the traditional matching methods, we propose to learn variable weights that remove the non-confounders and differentiate the importance of confounders during the matching process.

To be specific, for each treated unit i, we find its closest match among control units as follows:

$$match(i) = \underset{j:T_j=0}{\arg\min}\, d(X_i \odot \beta, X_j \odot \beta), \tag{7}$$

where $\odot$ refers to Hadamard product and $\beta \in \mathbb{R}^{p\times 1}$ is the variable weights to select and differentiate the confounders from all observed variables.

Next, we will give a theoretical analysis on how to learn the variable weights β for differentiated matching.

The general relationship among observed variables $\mathbf{X}$, treatment T and outcome Y can be represented as:

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon, \tag{8}$$

where the *true* ITT_i for a treated units i is $g(\mathbf{X}_i)$, and the potential outcome $Y(0)$ can be represented by:

$$Y(0) = f(\mathbf{X}) + \epsilon. \tag{9}$$

Proposition 1 *In matching, the bias of the treatment effect is mainly caused by the difference in the potential outcome $Y(0)$ between the matched treated and the control units.*

For each treated unit i and its matched control unit j, with Eq. (2) and Eq. (8), we have

$$\begin{aligned}\widehat{ITT}_i &= Y_i^{obs} - Y_j^{obs}\\ &= [f(X_i) + g(X_i) + \epsilon_i] - [f(X_j) + \epsilon_j]\\ &= g(\mathbf{X}_i) + (f(X_i) + \epsilon_i - f(X_j) - \epsilon_j)\\ &= ITT_i + [Y_i(0) - Y_j(0)]\end{aligned}$$

where $[Y_i(0) - Y_j(0)]$ refers to the difference of potential outcome $Y(0)$ between the matched treated-control pair (i, j).

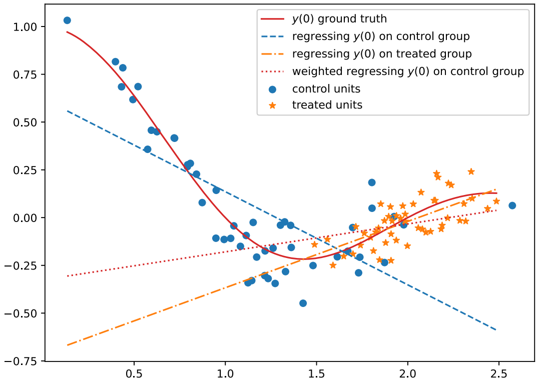


Fig. 1 Demonstration on distribution shift between treated and control groups

Proposition 2 *In matching, not all observed variables are confounders, and different confounders make unequal confounding bias on treatment effect with their own weights. The weights can be learned via regressing potential outcome* $Y(0)$ *on observed variables* $\mathbf{X}$.

We prove Proposition 2 with following assumption.

Assumption 1. The regression of potential outcome $Y(0)$ on observed variables $\mathbf{X}$ is linear,[1] that is $f(\mathbf{X}) = c + \mathbf{X}\beta$.

Under assumption 1, the bias of ITT_i can be rewritten as:

$$\begin{aligned}[Y_i(0) - Y_j(0)] &= (c + X_i\beta + \epsilon_i) - (c + X_j\beta + \epsilon_j)\\ &= (X_i\beta - X_j\beta) + \phi(\epsilon)\\ &= \sum_{k=1}^{p} \beta_k(X_{i,k} - X_{j,k}) + \phi(\epsilon).\end{aligned}$$

where $\phi(\epsilon) = \epsilon_i - \epsilon_j$ refers to the difference of noises between the matched treated and control unit pair. In order to reduce the bias of estimated ITT, we have to minimize the term $\sum_{k=1}^{p} \beta_k(X_{i,k} - X_{j,k})$, where $(X_{i,k} - X_{j,k})$ means the difference of the k^{th} variable between treated and control units, and the parameter β_k represents the confounding bias weight of the k^{th} variable.

Therefore, different confounders make unequal confounding bias with their own weights, and those variables with weight $\beta_k = 0$ are non-confounders since they would not bring any bias. Fortunatelly, the confounder weight β_k is exactly the coefficient of X_k in the function $f(\mathbf{X})$. Hence, we can learn the confounder weights from the regression of potential outcome $Y(0)$ on observed variables $\mathbf{X}$.

4.3 Confounder weights learning

From proposition 2, we know that the confounder weights can be learned via regressing potential outcome $Y(0)$ on observed variables $\mathbf{X}$. However, in observational data, we

[1] The linear assumption can be relaxed by adding high order terms in the regression process.

can only observe the potential outcome $Y(0)$ from the control units with $T = 0$. When facing model misspecification, direct regression fails dual to distribution shift between treated and control groups (Austin 2011; Li et al. 2020).

Figure 1 shows an example to illustrate the distribution shift problem. In the example, data points lie on $y = \frac{sin \pi x}{\pi x}$ with random noise sampled from normal distribution $\mathcal{N}(0, (0.1)^2)$. The control units and treated units are generated from two distribution $\mathcal{N}\left(0, \left(\frac{1}{2}\right)^2\right)$ and $\mathcal{N}\left(0, \left(\frac{1}{4}\right)^2\right)$.

Then, the regression model tries to learn the confounder weights β, which minimizes the expected value of the following loss function over the distribution of control units (T=0):

$$\beta = \arg\min_{\beta} \mathbf{E}_{X,Y,T \sim D}[l(f(X, \beta), Y)|T = 0], \tag{10}$$

where D refers to observed data. The regression ends up with the blue dashed line in Fig. 1.

However, as defined in Eqs. 2 and 3, we are interested in the treatment effect on the treated units. Hence, an ideal way to learn the confounder weights should be based on the potential outcome $Y(0)$ from treated units (the orange dash-dot line in Fig. 1). The presence of confounders would bring covariate/distribution shift between treated ($T = 1$) and control ($T = 0$) units, leading to the distribution shift problem in learning confounder weights in Eq. 10. Inspired by the techniques from covariate shift (Zadrozny 2004), we propose to adjust the distribution of control units to mimic one of the treated units via unit reweighting with the following proposition.

Proposition 3 *Given the distribution $P(\mathbf{X})$ of treated units $D_{T=1}$ and control units $D_{T=0}$, the distribution shift problem between $D_{T=1}$ and $D_{T=0}$ can be addressed by reweighting control units with sample weights $W = \frac{P_D(T=0)}{P_D(T=1)} \cdot \frac{P_D(T=1|\mathbf{X})}{P_D(T=0|\mathbf{X})}$.*

Proof

$$\begin{aligned} W \cdot P_{D_{T=0}}(\mathbf{X}) &= W \cdot P_D(\mathbf{X}|T = 0) \\ &= \frac{P_D(T = 0)}{P_D(T = 1)} \cdot \frac{P_D(T = 1|\mathbf{X})}{P_D(T = 0|\mathbf{X})} \cdot \frac{P_D(T = 0|\mathbf{X}) P_D(\mathbf{X})}{P_D(T = 0)} \\ &= \frac{P_D(T = 1|\mathbf{X}) P_D(X)}{P_D(T = 1)} \\ &= P_D(\mathbf{X}|T = 1) = P_{D_{T=1}}(\mathbf{X}) \end{aligned}$$

□

Hence, to address the distribution shift problem between treated and control units for confounder weights learning, one should reweight the control units with the sample weights W in proposition 3. Then, the confounder weights are learned as follows:

$$\beta = \arg\min_{\beta} \mathbf{E}_{X,Y,T \sim D}[W \cdot l(f(X, \beta), Y)|T = 0]. \tag{11}$$

Through the weighted regression, the learned line is shown as the red dotted line in Fig. 1, which is more close to the desired results.

5 Algorithm

5.1 Differentiated matching

Based on proposition 2, assumption 4 and Eq. 10, we first propose to learn the variable weights β for confounder selection and differentiation before units matching as follows:

$$\beta = \arg\min_{\beta} \sum_{j:T_j=0} (Y_j - X_j\beta)^2$$
$$s.t.\|\beta\|_1 \leq \lambda \quad and \quad \|\beta\|_2^2 \leq \delta, \tag{12}$$

where β is the confounder weights. With the constraints $\|\beta\|_1 \leq \lambda$ and $\|\beta\|_2^2 \leq \delta$, we can remove the non-confounders and smooth the confounder weights.

With the variable weights β learned from Eq. 12, we propose to match units with Differentiated Matching (DM) algorithm combined with Eq. (6). Specifically, we adopt absolute distance to measure the similarity between two units, then, for each treated unit i, we match it to a control unit:

$$match(i) = \arg\min_{j:T_j=0} |\beta \cdot (X_i - X_j)|. \tag{13}$$

Then, for each treated unit i, we obtain a matched unit pair (i, j), where $j = match(i)$. To bound the bias during matching process, we drop unit pair if $|\beta \cdot (X_i - X_j)| > \epsilon$. Finally, we obtain a matched treated and control unit pairs set $S = \{(i, j)|T_i = 1, T_j = 0\}$, and one can easily estimate individual treatment effect ITT_i for each treated unit i with Eq. (4) and average treatment effect ATT with Eq. (5).

5.2 Propensity score based DM

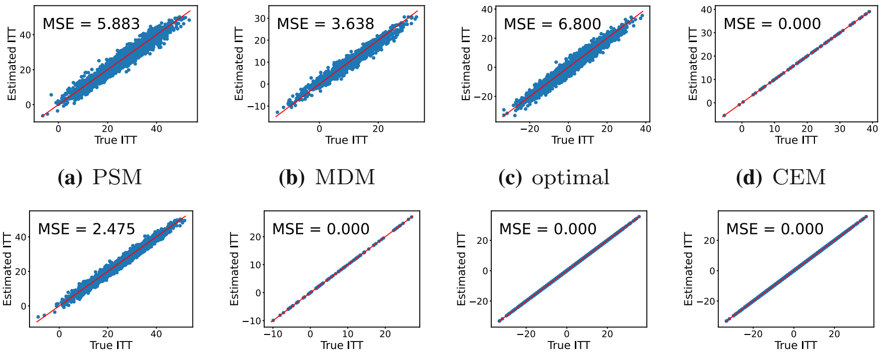
To address the distribution shift problem between treated and control units with the proposition 3, we further propose to learn confounder weights β based on Eq. 11.

The sample weights W in proposition 3 can be decomposed into two components, including $\frac{P_D(T=0)}{P_D(T=1))}$, which is a constant for each unit, and $\frac{P_D(T=1|\mathbf{X})}{P_D(T=0|\mathbf{X})}$, which can be estimated by the propensity score $e(X) = P_D(T = 1|X)$. Hence, we propose the following propensity score weighted regression for learning the confounder weights

Algorithm 1 (Propensity Score based) Differentiated Matching Algorithm

Input: Observed variable matrix **X**, Treatment variable T, Outcome Y, estimated propensity score $\hat{e}$, tradeoff parameters $\lambda > 0$, $\delta > 0$ and a predefined threshold ϵ
Output: ITT_i for each treated unit i, ATT

```
Optimize confounder weight $\beta$ with Eq. 12 or 14
Constructing Ball Tree with control units $X, T = 0$.
Obtain $S = \{(i, j)|T_i = 1, T_j = 0\}$ by matching units based on Eq. (13)
Calculate the $ITT_i$ for each treated-control unit pair $(i, j)$ with Eq. (4)
Calculate the $ATT$ on all matched units pairs for each treated unit $i$ based on its match with Eq. (5)
return $ITT_i$, $ATT$
```



(a) PSM **(b)** MDM **(c)** optimal **(d)** CEM

(e) genMatch **(f)** DAME **(g)** DM **(h)** PSDM

Fig. 2 Estimated ITT v.s. true ITT in categorical setting with $n = 5000$, $p = 15$. Each blue dot represents the results of ITT estimation on a treated unit. The red line $y = x$ represents the estimated ITT exactly equals the true ITT

β

$$\beta = \arg\min_{\beta} \sum_{j:T_j=0} \frac{\hat{e}(X_j)}{1-\hat{e}(X_j)} (Y_j - X_j\beta)^2$$
$$s.t. \|\beta\|_1 \leq \lambda \quad and \quad \|\beta\|_2^2 \leq \delta, \tag{14}$$

where the propensity score $\hat{e}$ can be estimated with a Logistic regression on treatment via learning $P_D(T = 1|X)$.

Based on the confounder weights learned from Eq. 14, we can also estimate the ITT and ATT by matching with Eq. 13, and we name this algorithm as Propensity Score based Differentiated Matching (PSDM).

The details of the proposed (propensity score based) differentiated matching algorithm are described in Algorithm 1.

In experiments, we propose to adopt a data structure named ball tree (Omohundro 1989) during unit matching, which can efficiently reduce the complexity of matching from $O(n^2p)$ to $O(nlog(n)p)$, where n refers to sample size and p is the dimension of variables.

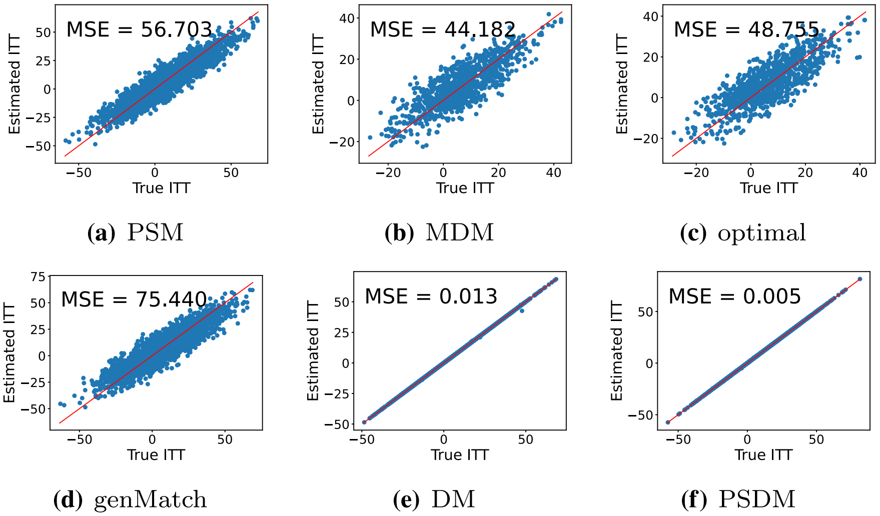


Fig. 3 Estimated ITT v.s. true ITT in continuous setting with $n = 5000$, $p = 50$. Each blue dot represents the ITT estimation of a treated unit. The red line $y = x$ represents the estimated ITT exactly equals the true ITT. (DAME cannot be applied for continuous variables.)

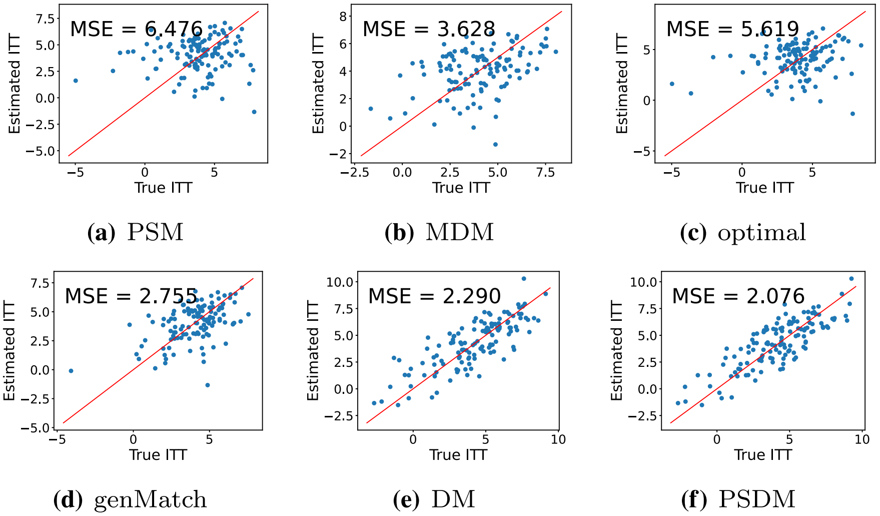


Fig. 4 Estimated ITT v.s. true ITT of IHDP dataset. Each blue dot represents the results of a treated unit and the red line $y = x$ represents the estimated ITT exactly equals the true ITT

6 Experiments

In this section, we apply our algorithm on both synthetic and real-world datasets to demonstrate the advantages of our algorithm for treatment effect estimation.

6.1 Baselines

In this work, we focus on the interpretable matching methods for treatment effect (ITT and ATT) estimation, so we implement the following baselines for comparison:

- Naive estimator, which evaluates the ATT by directly comparing the average outcome between the treated and control units, ignoring the confounding bias in data;
- Propensity Score Matching (PSM) (Austin 2011), which performs nearest neighbor matching on propensity score;
- Optimal Pair Match(optimal) (Rosenbaum 2017) tries to solve an optimization problem to form matches based on pre-defined distance measure;
- Mahalanobis Distance Matching(MDM) (Rosenbaum and Rubin 1985), which performs nearest neighbor matching based on Mahalanobis metric;
- Genetic Matching (genMatch) (Diamond and Sekhon 2013), which matches units based on the weighted Mahalanobis metric, where the weight is learned by evolutionary search;
- Dynamic Almost Matching Exactly matching (DAME) (Liu et al. 2019), which matches units based on the weighted Hamming distance and uses a dynamic program algorithm to make a selection of the confounders and irrelevant variables.

6.2 Evaluation metrics

To evaluate the performance of our proposed method, we carry out the experiments for 50 times independently. Based on the estimated ATT ($\widehat{ATT}$), we calculate its $Bias$, standard deviations (SD), mean absolute errors (MAE) and root mean square errors ($RMSE$) with following definitions:

$$Bias = |\frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k - ATT|$$

$$SD = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - \frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k)^2}$$

$$MAE = \frac{1}{K}\sum_{k=1}^{K}|\widehat{ATT}_k - ATT|$$

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - ATT)^2}$$

where K is the experimental times, $\widehat{ATT}_k$ is the estimated ATT in k^{th} experiment and ATT represents the *true treatment effect*.

6.3 Experiments on synthetic data

6.3.1 Datasets

We generate the synthetic datasets with different settings. **Categorical Setting:** In this setting, we set the observed variables as categorical. Specifically, we consider two sample sizes $n = \{2000, 5000\}$ and also vary the dimension of observed variables $p = \{10, 15, 20\}$.[2]

To bring the confounding bias, we generate the observed variables $X = \{x_1, x_2, \cdots, x_p\}$ with independent Gaussian distributions as:

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p \overset{iid}{\sim} \mathcal{N}(0.5, 2), \text{ for treated units.}$$
$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p \overset{iid}{\sim} \mathcal{N}(-0.5, 2), \text{ for control units.}$$

To make X categorical, we let $x_i = 1$ if $x_i \geq 0$, otherwise $x_i = 0$.
Continuous setting: Follow (Kuang et al. 2019), in continuous setting, we also consider two sample sizes $n = \{2000, 5000\}$ but vary the dimension of observed variables $p = \{10, 50, 100\}$. We generate the observed variables $\mathbf{X} = \{x_1, x_2, \ldots, x_p\}$ with independent Gaussian distributions as:

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p \overset{iid}{\sim} \mathcal{N}(0, 1). \tag{15}$$

Then, generating the binary treatment T as the following function of variable X to introduce confounding bias:

$$T \sim Bernoulli(1/(1 + exp(-\sum_{i=1}^{p \cdot r_c} s_c \cdot x_i + \mathcal{N}(0, 1)))). \tag{16}$$

Finally, the outcome Y is generated as:

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon, \tag{17}$$

where $f(\mathbf{X}) = \mathbf{X}\alpha$, $g(\mathbf{X}) = \mathbf{X}\gamma$, and the coefficients $\alpha, \gamma \in [0, 1]$ are randomly generated.

We also generated a nonlinear outcome with $f(X) = X\alpha + \beta \sum_{i,j\ j>i} x_i x_j$ to test the robustness of our model facing model misspecification. In synthetic data, we know the *true* ITT_i for each treated unit i and ATT over the whole dataset. We evaluate the ITT_i and ATT with our algorithm, comparing with baselines.

6.3.2 Results

We evaluate the performance of our proposed method on both ITT and ATT estimation.

[2] Higher dimension brings NULL matching in DAME and CEM, we omitted these methods in continuous settings.

Table 2 Results of ATT estimation in different settings. The “NA” represents the number of matched units is $zero$. We did not compare with DAME and CEM in continuous setting since DAME cannot be applied in continuous setting, and CEM brings NULL matching best performance is marked bold

Categorical setting										
	p	$p = 10$			$p = 15$			$p = 20$		
n	Estimator	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE
$n = 2000$	$\widehat{ATT}_{naive}$	13.8213(8.812)	14.7573	17.188	18.6837(10.424)	19.0561	21.721	27.2493(12.543)	27.2631	30.010
	$\widehat{ATT}_{CEM}$	1.4142(1.148)	1.5258	1.822	2.5697(2.980)	3.1382	3.935	NA(NA)	NA	NA
	$\widehat{ATT}_{PSM}$	0.0203(0.433)	0.3402	0.434	0.0951(0.396)	0.3296	0.407	0.0895(0.535)	0.4103	0.542
	$\widehat{ATT}_{optimal}$	0.0409(0.403)	0.3154	0.405	0.1132(0.368)	0.3132	0.385	0.0191(0.465)	0.3619	0.465
	$\widehat{ATT}_{MDM}$	0.0319(0.559)	0.4800	0.560	0.0076(0.843)	0.6888	0.843	0.0112(0.985)	0.7703	0.985
	$\widehat{ATT}_{GenMatch}$	0.5825(0.586)	0.6820	0.827	0.0071(0.376)	0.3040	0.376	0.0111(0.484)	0.3895	0.484
	$\widehat{ATT}_{DAME}$	0.0541(0.127)	0.1433	0.191	0.1006(0.269)	0.3645	0.453	NA(NA)	NA	NA
	$\widehat{ATT}_{DM}$	0.0119(0.042)	**0.0328**	**0.053**	0.0306(0.112)	0.0880	0.143	0.0016(0.122)	0.0849	0.148
	$\widehat{ATT}_{PSDM}$	**0.0078(0.041)**	0.0344	0.054	**0.0272(0.112)**	**0.0876**	**0.142**	**0.0003(0.113)**	**0.0760**	**0.136**
$n = 5000$	$\widehat{ATT}_{naive}$	13.8779(9.425)	14.2151	17.056	17.4313(9.485)	17.7010	20.082	27.4486(14.434)	27.5014	31.059
	$\widehat{ATT}_{CEM}$	0.9256(0.833)	1.0565	1.245	2.6574(1.878)	2.8008	3.254	NA(NA)	NA	NA
	$\widehat{ATT}_{PSM}$	0.0760(0.494)	0.4081	0.500	0.0489(0.605)	0.5125	0.607	0.0675(0.552)	0.4522	0.556
	$\widehat{ATT}_{optimal}$	0.0950(0.475)	0.3922	0.484	0.0268(0.599)	0.5066	0.600	0.0895(0.535)	0.4428	0.543
	$\widehat{ATT}_{MDM}$	0.1106(0.564)	0.4507	0.575	0.0640(0.787)	0.6279	0.790	0.0299(0.747)	0.5893	0.748
	$\widehat{ATT}_{GenMatch}$	1.3127(1.001)	1.3469	1.651	0.1140(0.322)	0.2844	0.342	0.0195(0.350)	0.3050	0.351
	$\widehat{ATT}_{DAME}$	0.0714(0.107)	0.1106	0.154	0.0796(0.216)	0.2175	0.307	NA(NA)	NA	NA
	$\widehat{ATT}_{DM}$	**0.0030(0.031)**	0.0141	0.034	0.0036(0.024)	**0.0235**	0.034	**0.0002(0.036)**	**0.0311**	**0.047**
	$\widehat{ATT}_{PSDM}$	0.0033(0.031)	**0.0133**	**0.033**	**0.0034(0.024)**	0.0245	**0.034**	0.0017(0.039)	0.0357	0.053

Table 2 continued

Categorical setting										
	p	$p = 10$			$p = 15$			$p = 20$		
n	Estimator	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE
Continuous setting										
	p	$p = 20$			$p = 50$			$p = 100$		
n	Estimator	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE
$n = 2000$	$\widehat{ATT}_{naive}$	4.2753(1.386)	4.2753	4.494	6.5867(1.636)	6.5867	6.787	9.0526(1.593)	9.0526	9.192
	$\widehat{ATT}_{PSM}$	0.4892(1.422)	1.1953	1.504	0.3556(1.620)	1.3764	1.658	0.7678(1.828)	1.5769	1.982
	$\widehat{ATT}_{optimal}$	0.4338(1.420)	1.1396	1.485	0.3004(1.681)	1.4018	1.708	0.6522(1.753)	1.4782	1.870
	$\widehat{ATT}_{MDM}$	0.4390(1.404)	1.1467	1.471	0.2668(1.658)	1.4007	1.679	0.5666(1.737)	1.4575	1.827
	$\widehat{ATT}_{GenMatch}$	0.2927(1.104)	0.9348	1.142	0.2801(1.512)	1.2574	1.538	0.5953(1.628)	1.4035	1.734
	$\widehat{ATT}_{DM}$	0.0028(0.022)	0.0144	0.026	0.0053(0.020)	0.0208	0.029	0.0042(0.033)	0.0350	0.048
	$\widehat{ATT}_{PSDM}$	**0.0010(0.010)**	**0.0064**	**0.012**	**0.0023(0.007)**	**0.0058**	**0.009**	**0.0022(0.008)**	**0.0072**	**0.011**
$n = 5000$	$\widehat{ATT}_{naive}$	3.9457(1.611)	3.9983	4.311	6.4285(1.635)	6.4285	6.633	8.5796(1.718)	8.5796	8.750
	$\widehat{ATT}_{PSM}$	0.0207(1.461)	1.1526	1.461	0.0184(1.578)	1.3548	1.578	0.1871(1.607)	1.1624	1.618
	$\widehat{ATT}_{optimal}$	0.0610(1.467)	1.1716	1.469	0.0211(1.605)	1.3562	1.606	0.1641(1.537)	1.1451	1.545
	$\widehat{ATT}_{MDM}$	0.0301(1.455)	1.1594	1.456	0.0203(1.626)	1.3904	1.626	0.2109(1.636)	1.1753	1.650
	$\widehat{ATT}_{GenMatch}$	0.0881(1.016)	0.8399	1.020	0.0814(1.346)	1.1050	1.349	0.1479(1.418)	1.0591	1.425
	$\widehat{ATT}_{DM}$	**0.0003(0.008)**	0.0057	0.010	0.0005(0.011)	0.0082	0.014	0.0029(0.013)	0.0137	0.019
	$\widehat{ATT}_{PSDM}$	0.0006(0.004)	**0.0021**	**0.004**	**0.0001(0.003)**	**0.0019**	**0.003**	**0.0004(0.003)**	**0.0026**	**0.004**

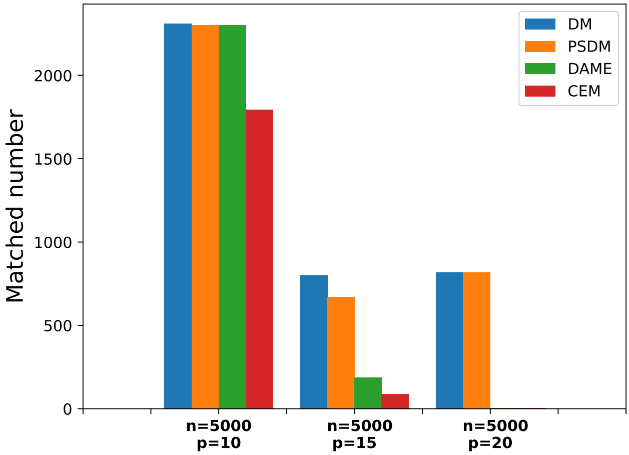


Fig. 5 No. of matched unit pairs in categorical setting

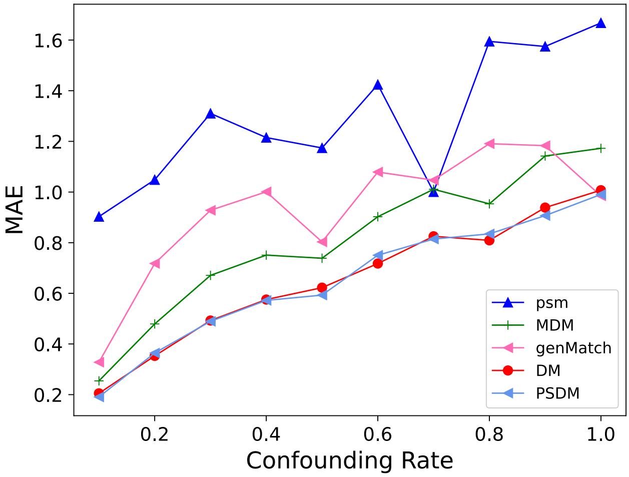
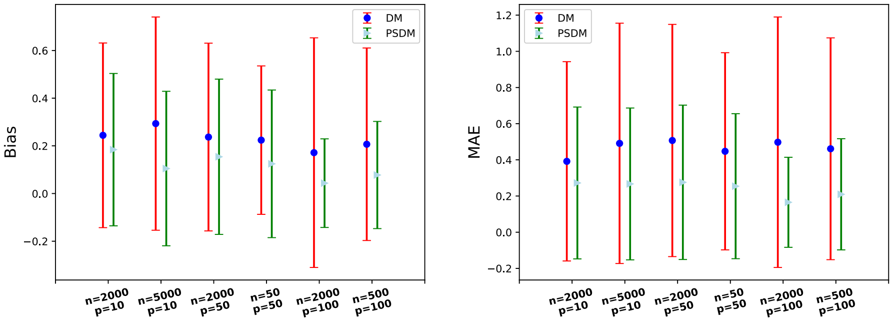


Fig. 6 Experiment explores different confounding rate

Figures 2 and 3 scatter plots the performance on ITT estimation in categorical and continuous settings. The MSE refers to the mean square error between the estimated and true ITT over all matched units (Fig. 4).

DAME and CEM achieve good performance in the categorical setting. However, these two methods are not suitable to be extended into continuous and high-dimensional settings. To demonstrate the advantages of our methods over DAME and CEM, we also compared the number of matched unit pairs in Fig. 5, which shows that our method can consistently obtain more high-quality matched units than DAME and CEM under all settings, especially in settings with higher dimension of variables. From Figs. 2 and 3, we conclude that our algorithms have a significant improvement over other methods on ITT estimation (Fig. 6).



(a) Bias

(b) MAE

Fig. 7 Nonlinear experiment with different settings

Table 2 demonstrates the results on ATT estimation in different settings, where the $Bias$ refers to the absolute error between the true and estimated ATT. The SD, MAE and $RMSE$ represent the standard deviations, mean absolute errors and root mean square errors of estimated ATT ($\widehat{ATT}$) after 50 times independently experiments. Form Table 2, we have the following observations and analyses: (1) Most matching methods do not distinguish the importance of confounders, and match units based on all variables. Hence, these matching methods have poor performance in ATT estimation. (2) DAME considers the selection of confounders, but ignores confounder differentiation and distribution shift problem between treated and control units, leading to underperformance in estimating ATT. (3) With considering confounder differentiation and distribution shift problem in confounder weights learning, our DM and PSDM methods achieved a significant improvement in treatment effect estimation.

Figure 7 compares our DM and PSDM methods in the non-linear setting, where the regression model is misspecified but the propensity score model is correctly specified. In this setting, weighted regression by propensity score in PSDM can help to improve the effectiveness (smaller MAE) and robustness (smaller standard deviation) than DM, especially in the high dimensional settings.

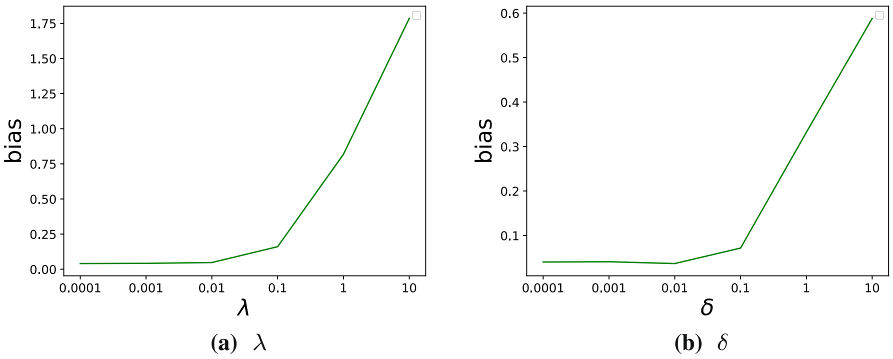
Figure 6 explores the models' performance with the increase of confounding rate s_c in Eq. 16, which measures the confounding bias of the generated dataset. The results show that as the confounding rate increases, our algorithm is very stable compared to other methods.

6.3.3 Parameter analysis

The parameter λ and δ in Eq. 14 is related to the regularization in the supervised regression. These parameters are selected by grid searching, each parameter is varied from $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. The result is shown in Fig. 8, which shows the treatment effect estimation $Bias$ with different λ and δ. From the result, we know that the performance of our model is relatively stable when $\lambda, \delta < 0.01$. At the same time, when $\lambda = 0.0001$ and $\delta = 0.01$, it reaches the best performance.

Table 3 Results of ATT estimation in different settings. The “NA” represents the number of matched units is $zero$. We did not compare with DAME and CEM in continuous setting since these methods cannot be applied in continuous setting

Continuous setting										
	p	$p = 20$			$p = 50$			$p = 100$		
n	Estimator	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE	Bias(SD)	MAE	RMSE
$n = 2000$	$\widehat{ATT}_{naive}$	0.2847(0.840)	1.1391	1.415	0.2344(0.919)	1.2593	1.559	0.2558(0.992)	1.3696	1.691
	$\widehat{ATT}_{PSM}$	1.6466(2.719)	2.5779	3.179	5.7322(6.728)	6.9498	8.839	7.6299(4.984)	8.0256	9.114
	$\widehat{ATT}_{optimal}$	0.3041(1.385)	1.1296	1.418	0.2204(1.736)	1.4249	1.750	0.0976(1.751)	1.4332	1.754
	$\widehat{ATT}_{MDM}$	0.3368(1.387)	1.1352	1.427	0.2900(1.715)	1.3885	1.739	0.2772(1.714)	1.3642	1.737
	$\widehat{ATT}_{GenMatch}$	0.2108(1.168)	0.9481	1.187	0.2788(1.363)	1.1344	1.391	0.1652(1.557)	1.2943	1.566
	$\widehat{ATT}_{DM}$	0.0018(0.006)	0.0045	0.007	0.0004(0.005)	0.0056	0.008	0.0035(0.009)	0.0077	0.012
	$\widehat{ATT}_{PSDM}$	0.0018(0.006)	0.0045	0.007	0.0004(0.005)	0.0056	0.008	0.0035(0.009)	0.0077	0.012
$n = 5000$	$\widehat{ATT}_{naive}$	0.1939(0.916)	1.1917	1.503	0.0239(0.896)	1.2696	1.554	0.0980(1.070)	1.5398	1.875
	$\widehat{ATT}_{PSM}$	1.5652(2.767)	2.5732	3.179	4.1828(4.230)	4.6524	5.949	7.8867(7.536)	8.7234	10.908
	$\widehat{ATT}_{optimal}$	0.1726(1.516)	1.2108	1.526	0.1030(1.585)	1.2646	1.588	0.1281(1.848)	1.4955	1.853
	$\widehat{ATT}_{MDM}$	0.2103(1.511)	1.2272	1.526	0.0462(1.573)	1.2772	1.574	0.0553(1.819)	1.4459	1.820
	$\widehat{ATT}_{GenMatch}$	0.0534(1.184)	0.9454	1.185	0.0026(1.381)	1.1061	1.381	0.0779(1.674)	1.4120	1.676
	$\widehat{ATT}_{DM}$	0.0008(0.007)	0.0025	0.007	0.0002(0.003)	0.0024	0.004	0.0004(0.004)	0.0032	0.005
	$\widehat{ATT}_{PSDM}$	0.0008(0.007)	0.0025	0.007	0.0002(0.003)	0.0024	0.004	0.0004(0.004)	0.0032	0.005



(a) λ **(b)** δ

Fig. 8 Hyper-parameter sensitivity analysis of λ and δ

6.3.4 Performance of confounder selection

We also compared our algorithm with other baselines in the ability of handling non-confounders. We consider two sample sizes $n = \{2000, 5000\}$ but vary in the dimension of observed variables $p = \{10, 50, 100\}$.

And we generate the observed variables $\mathbf{X} = \{x_1, x_2, \ldots, x_p, x_1^{non}, x_2^{non}, \cdots, x_{p_{non}}^{non}\}$ with independent Gaussian distributions as:

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p, \mathbf{x}_1^{non}, \mathbf{x}_2^{non}, \cdots, \mathbf{x}_{p_{non}}^{non} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{18}$$

where $p_{non} = p * r_n$. The binary treatment variable T is generated as a function of variable X to introduce confounding bias from a logistic function with Eq. (16).

Finally the outcome Y is generated as:

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon, \tag{19}$$

where $f(\mathbf{X}) = \mathbf{X}\alpha$, $g(\mathbf{X}) = \mathbf{X}\gamma$, and the coefficients $\alpha_{1:p}, \gamma_{1:P} \in [0, 1]$ are randomly generated and $\alpha_{p+1:p+p_{non}}, \gamma_{p+1:p+p_{non}} \in \{0\}$. We also set $r_n = 0.25$.

The results in Table 3 show that our algorithm performs well when handling data with non-confounders. Methods like PSM, Optimal matching and MDM may aggravate the confounding bias because these methods match units based on all observed confounders.

We also evaluate the performance of confounder selection. When facing linear settings, the regression model can do confounder selection perfectly, which means that learned variable weights of confounders satisfy that $\beta_i > \epsilon, w_i \neq 0$.

To fully explore the ability of our algorithm in confounder selection, we test our model's performance in non-linear settings, where our model is non-specified. The datasets vary in variables' dimension $p = \{10, 15, 20, 30, 50\}$. And the data generating process is similar to the previous setting with adding non-linear term $x_i x_j,\ j > i$. Each experiment is performed independently 50 times.

We discuss the precision of the confounder selection, each confounder x_i is precisely selected means that $\beta_i > \epsilon$. In experiments, we set $\epsilon = 10^{-5}$. The result is shown

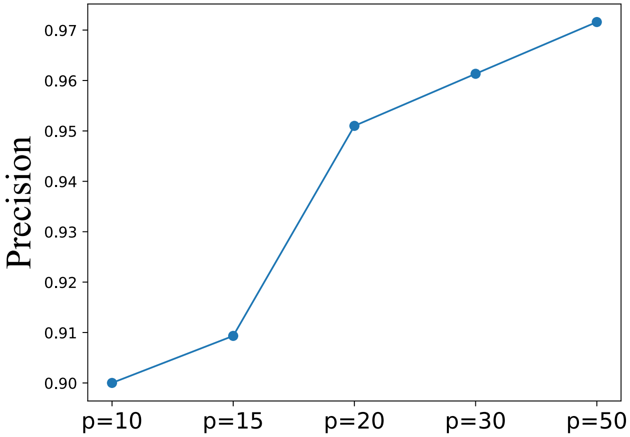


Fig. 9 Results of confounder selection in non-linear settings

in Fig. 9. From this figure, we conclude that our models still have good performance doing confounder selection even facing model misspecification.

6.4 Experiments on real world data

6.4.1 Datasets

We compare all estimators on the following datasets:
LaLonde dataset LaLonde (1986) is an authoritative dataset in the field of causal inference. We adopted the dataset as Dehejia and Wahba (1999) did and replaced the control group with the control group obtained by PSID (Population Survey of Income Dynamics) sampling. We evaluate the ATT on the LaLonde dataset.
IHDP dataset Hill (2011) is a semi-synthetic dataset based on the Infant Health and Development Program(IHDP), which aims to measure the effect of home visits by specialists on future cognitive scores. We adopted the same dataset as Shalit et al. (2017) did and compared our algorithm with other baseline methods on both ITT and ATT estimation.

6.4.2 Results

The result of ATT estimation of Lalonde dataset is plotted in Fig. 10. Due to the existence of confounding bias, directly compare the treated and control group even leads to a huge bias on ATT estimation. The other baseline methods effectively reduce confounding bias, however not good enough compared with our methods. Our DM and PSDM methods achieve the best performance on ATT estimation.

Table 4 and Fig. 4 demonstrate the estimation of ATT and ITT on IHDP dataset. Table 4 shows that our algorithm outperforms other baseline matching methods in estimating ATT. Moreover, integrate propensity score in the process of confounder weights learning, our PSDM achieves a more accurate and robust estimation on ATT.

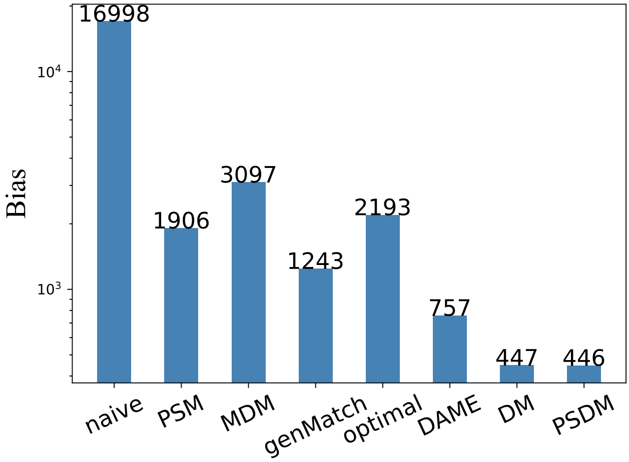


Fig. 10 Bias of ATT estimation on LaLonde dataset

Table 4 Results of ATT estimation of IHDP dataset, the best performance is marked bold

Estimator	Bias(SD)	MAE	RMSE
$\widehat{ATT}_{naive}$	0.6722(1.906)	1.0754	2.189
$\widehat{ATT}_{PSM}$	0.2112(1.125)	0.5390	1.145
$\widehat{ATT}_{optimal}$	0.5701(2.223)	1.1326	2.295
$\widehat{ATT}_{MDM}$	0.3938(1.180)	0.5340	1.244
$\widehat{ATT}_{GenMatch}$	0.2693(1.052)	0.4385	1.086
$\widehat{ATT}_{DM}$	0.0838(0.383)	0.2360	0.450
$\widehat{ATT}_{PSDM}$	**0.0335(0.229)**	**0.1841**	**0.294**

Figure 4 clearly demonstrate the advantages of our proposed algorithms on ITT estimation. With considering confounder differentiation and distribution shift on confounder weights learning, our algorithms (both DM and PSDM) achieved smaller MSE on ITT estimation.

7 Conclusion

In this work, we focus on estimating both individual and average treatment effects more precisely by matching methods. We argue that traditional matching methods do not consider the difference among confounders, leading to underperformance in treatment effect estimation. We proposed a differentiated matching algorithm with a theoretical guarantee for treatment effect estimation. A propensity score based differentiated matching algorithm is also proposed to address the distribution shift problem in confounder weights learning. Extensive experiments on both synthetic and real-world datasets demonstrate that our proposed algorithms outperform state-of-the-art matching methods on individual and average treatment effect estimation.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (No. 62006207, No. 62037001), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), Key R & D Projects of the Ministry of Science and Technology (2020YFC0832500), Project by Shanghai AI Laboratory (P22KS00111), Key Laboratory for Corneal Diseases Research of Zhejiang Province and the Fundamental Research Funds for the Central Universities (226-2022-00142).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46(3):399–424

Bottou L, Peters J, Quiñonero-Candela J, Charles DX, Chickering DM, Portugaly E, Ray D, Simard P, Snelson E (2013) Counterfactual reasoning and learning systems: the example of computational advertising. J Mach Learn Res 14(1):3207–3260

Chan D, Ge R, Gershony O, Hesterberg T, Lambert D (2010) Evaluating online ad campaigns in a pipeline: causal models at scale. In: KDD, pp 7–16

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94(448):1053–1062

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95(3):932–945

Hill JL (2011) Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 20(1):217–240

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: coarsened exact matching. Polit Anal 20(1):1–24

Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, Cambridge

Kallus N (2017) A framework for optimal matching for causal inference. In: Artificial Intelligence and Statistics, pp 372–381

Kallus N (2019) Generalized optimal matching methods for causal inference. J Mach Learn Res (forthcoming)

Kohavi R, Longbotham R (2011) Unexpected results in online controlled experiments. ACM SIGKDD Explor Newsl 12(2):31–35

Kuang K, Cui P, Li B, Jiang M, Wang Y, Wu F, Yang S (2019) Treatment effect estimation via differentiated confounder balancing and regression. ACM Trans Knowledge Dis from Data (TKDD) 14(1):1–25

Kuang K, Li L, Geng Z, Xu L, Zhang K, Liao B, Huang H, Ding P, Miao W, Jiang Z (2020) Causal inference. Engineering 6(3):253–263

LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econom Rev pp 604–620

Lewis RA, Reiley D (2008) Does retail advertising work? measuring the effects of advertising on sales via a controlled experiment on yahoo! Measuring the Effects of Advertising on Sales Via a Controlled Experiment on Yahoo

Li Y, Kuang K, Li B, Cui P, Tao J, Yang H, Wu F (2020) Continuous treatment effect estimation via generative adversarial de-confounding. In: Proceedings of the 2020 KDD Workshop on Causal Discovery, PMLR, pp 4–22

Liu Y, Dieng A, Roy S, Rudin C, Volfovsky A (2019) Interpretable almost matching exactly for causal inference. AISTATS

Omohundro SM (1989) Five balltree construction algorithms. Int Comput Sci Institute Berkeley

Rosenbaum PR (2017) Imposing minimax and quantile constraints on optimal matching in observational studies. J Comput Graph Stat 26(1):66–78

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33–38

Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: Int Conf Mach Learn, PMLR, pp 3076–3085

Wang T, Morucci M, Awan MU, Liu Y, Roy S, Rudin C, Volfovsky A (2021) Flame: A fast large-scale almost matching exactly approach to causal inference. J Mach Learn Res 22:1–41

Zadrozny B (2004) Learning and evaluating classifiers under sample selection bias. In: Proceedings of the twenty-first international conference on Machine learning, p 114

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Zhao Ziyu[1] · Kun Kuang[1] · Bo Li[2] · Peng Cui[2] · Runze Wu[3] · Jun Xiao[1] · Fei Wu[4,5,6]

Zhao Ziyu
benzhao.styx@gmail.com

Bo Li
libo@sem.tsinghua.edu.cn

Peng Cui
cuip@tsinghua.edu.cn

Runze Wu
wurunze1@corp.netease.com

Jun Xiao
junx@cs.zju.edu.cn

[1] Department of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang Province, China

[2] Tsinghua University, Beijing, China

[3] NetEase Fuxi AI Lab, Hangzhou, Zhejiang Province, China

[4] Institute of Artificial Intelligence, Zhejiang University, Hangzhou, China

[5] Shanghai Institute for Advanced Study of Zhejiang University, Shanghai, China

[6] Shanghai AI Laboratory, Shanghai, China