Data-Driven Variable Decomposition for Treatment Effect Estimation

Kun Kuang, Peng Cui, Hao Zou, Bo Li, Jianrong Tao, Fei Wu, and Shiqiang Yang

Abstract—Causal Inference plays an important role in decision making in many fields, such as social marketing, healthcare, and public policy. One fundamental problem in causal inference is the treatment effect estimation in observational studies when variables are confounded. Controlling for confounding effects is generally handled by propensity score. But it treats all observed variables as confounders and ignores the adjustment variables, which have no influence on treatment but are predictive of the outcome. Recently, it has been demonstrated that the adjustment variables are effective in reducing the variance of the estimated treatment effect. However, how to automatically separate the confounders and adjustment variables in observational studies is still an open problem, especially in the scenarios of high dimensional variables, which are common in the big data era. In this paper, we first propose a Data-Driven Variable Decomposition (D^2VD) algorithm, which can 1) automatically separate confounders and adjustment variables with a data-driven approach, and 2) simultaneously estimate treatment effect in observational studies with high dimensional variables. Under standard assumptions, we theoretically prove that our D^2VD algorithm can unbiased estimate treatment effect and achieve lower variance than traditional propensity score based methods. Moreover, to address the challenges from high-dimensional variables and nonlinear, we extend our D^2VD to a non-linear version, namely Nonlinear-D^2VD (N-D^2VD) algorithm. To validate the effectiveness of our proposed algorithms, we conduct extensive experiments on both synthetic and real-world datasets. The experimental results demonstrate that our D^2VD and N-D^2VD algorithms can automatically separate the variables precisely, and estimate treatment effect more accurately and with tighter confidence intervals than the state-of-the-art methods. We also demonstrated that the top-ranked features by our algorithm have the best prediction performance on an online advertising dataset.

Index Terms—Treatment effect estimation, variable decomposition, adjustment variables, confounder separation

1 Introduction

Causal inference [1], which refers to the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect [2], is a powerful statistical modeling tool for explanatory analysis. One fundamental problem in causal inference is the treatment effect estimation, and its key challenge is to remove the confounding bias induced by the different distributions of confounders between treated and control units. Taking Fig. 1 as an example, if a research attempt to assess the effect of a drug T on patients' recovery Y from population data where the drug usage was a patient's choice. The data shown that gender X affects a patient's choice of drug as well as his/her chances of recovery. In this scenario, gender X is a confounder that confounds the relation between treatment T (the choice of drug) and the outcome Y (the recovery of patients) since the distribution of gender would be different among patients's groups with different choices of drug. The gold standard approaches for removing confounding bias are randomized experiments, for example, A/B testing [3], where different treatments are randomly assigned to units.[1] However, the fully randomized experiments are usually extremely expensive [4] or sometimes even infeasible [5] in many scenarios. Hence it is highly demanding to develop automatic statistical approaches to infer treatment effect in observational studies.

In literature, Rosenbaum and Rubin [6] proposed a statistical framework for treatment effect estimation based on propensity score adjustment. Such a framework has been widely used in the observational causal study, including matching, stratification, inverse weighting and regression on propensity score [7], [8], [9]. The inverse propensity weighting is one of the most commonly used methods and has been part of a large family of causal models known as the marginal structural model [10], [11]. With a combination of inverse propensity weighting and regression, [12] proposed a doubly robust estimator. These methods have been widely used in various fields, including economics [13], epidemiology [14], health care [15], social science [16] and advertising [17].

The essence of these methods is to eliminate the confounding impact of confounders so that the precision of treatment effect estimation can be significantly improved.

- *Kun Kuang and Fei Wu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China. E-mail: kunkuang@zju.edu.cn, wufei@cs.zju.edu.cn.*
- *Peng Cui, Hao Zou, and Shiqiang Yang are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: {cuip, yangshq}@tsinghua.edu.cn, ahio@163.com.*
- *Bo Li is with the School of Economics and Management, Tsinghua University, Beijing 100084, China. E-mail: libo@sem.tsinghua.edu.cn.*
- *Jianrong Tao is with the Fuxi AI Lab, NetEase Inc, Hangzhou, Zhejiang 310000, China. E-mail: hztaojianrong@corp.netease.com.*

Manuscript received 13 Sept. 2019; revised 7 May 2020; accepted 11 May 2020. Date of publication 3 July 2020; date of current version 1 Apr. 2022. (Corresponding authors: Peng Cui, Bo Li, and Fei Wu.) Recommended for acceptance by X. Li. Digital Object Identifier no. 10.1109/TKDE.2020.3006898

1. Units represent the objects of treatment. For example, in an online advertising campaign, the units refer to the users in the campaign.

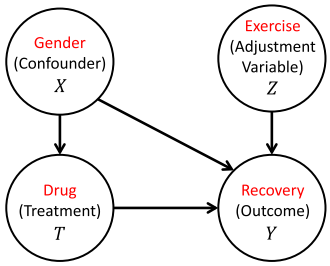


Fig. 1. An example of confounder, adjustment variable, treatment and outcome. The gender X is a confounder since it confounded the relation between the choice of drug T and patients' recovery Y. The exercise is a adjustment variable since it only influences the patients' recovery.

However, most of these works treat all observed variables as confounders when estimating the propensity score. Eventually, in the scenarios of high dimensional variables, some of them are not confounders but are predictive of the outcome, which are denoted by adjustment variables $\mathbf{Z}$ as shown in Fig. 1. For example, the exercise of a patient would not affect the choice of drug, but should influence the change of recovery, hence the exercise variable is not a confounder but a adjustment variable. Ignoring the adjustment variables will make the estimated treatment effect imprecise and with inflated variance.

Recently, some researchers have investigated the importance of the adjustment variables. [18], [19] have advocated that the adjustment variables should be included in the causal inference model. And [20] suggested that conditioning on such adjustment variables is unnecessary to remove bias but can reduce variance in treatment effect estimation. In a randomized experiment setting, [21] has proved that adjusting for the adjustment variables by lasso can reduce the variance of the estimated treatment effect.

All these methods in observational studies assume that the causal structure, i.e., whether a variable is the cause of the treatment or outcome, is known a priori. However, the causal structure cannot be well defined by prior knowledge in most cases, especially in the scenarios of high dimensional variables in the big data era. How to automatically separate confounders and adjustment variables in observational studies is still an open problem.

To address this problem, we propose a Data-Driven Variable Decomposition (D^2VD) algorithm to jointly optimize confounders' separation and Average Treatment Effect (ATE) estimation. More specifically, we propose a regularized integrated regression model, where a combined orthogonality and sparsity regularizer is constructed to simultaneously 1) separate the confounders and adjustment variables with a data-driven approach, 2) eliminate irrelevant variables which are neither confounders nor adjustment variables to avoid overfitting, and 3) estimate the ATE in observational studies. During estimating the ATE, the separated confounders can effectively eliminate their confounding impact on treatment, while the adjustment variables can significantly reduce the variances of estimated ATE through outcome adjustment. This enables us to estimate the true ATE more accurately and with tighter confidence intervals than baseline methods. Theoretically, we prove that our D^2VD algorithm can unbiasedly estimate the ATE and achieve lower estimation variance with considering the separation of confounders and adjustment variables. Moreover, to address the challenges from high-dimensional variables and nonlinear structure among variables, we adopt a neural network to learn a low dimensional and non-linear representation of variables and propose a Non-linear D^2VD (N-D^2VD) algorithm. With extensive experiments on both synthetic and real-world datasets, we demonstrate the effectiveness of our proposed algorithms on treatment effect estimation with observational data.

The main contributions in this paper are as follows:

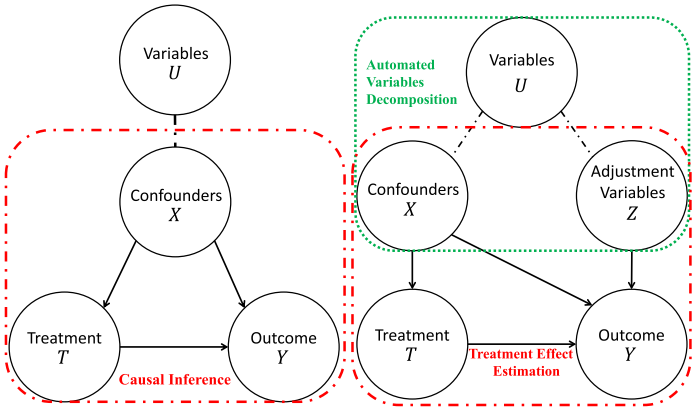
- We study a new problem of automatically separating confounders and adjustment variables, which is critical for the precision and confidence intervals of ATE estimation in observational studies.
- We propose a novel data-driven variables decomposition (D^2VD) algorithm, where a regularized integrated regression model is presented to enable confounder separation and ATE estimation simultaneously. Moreover, we extend our D^2VD algorithm to address the challenges from high-dimensional and non-linear and propose a Non-linear D^2VD algorithm.
- We give theoretical analysis on our proposed algorithm and prove that our algorithm can unbiasedly estimate the treatment effect with lower estimation variance by automatically confounder separation in observational studies.
- The advantages of our D^2VD and N-D^2VD algorithms are demonstrated in both synthetic and real-world data. It can also be straightforwardly applied to other causal inference studies, such as social marketing, health care, and public policy.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the adjusted estimator with considering variables decomposition. In Section 4, we theoretically give the bias analysis and variance analysis of our proposed estimation. Section 5 proposes the D^2VD algorithm under the linear assumption and gives its optimization that accurately infers the ATE. In Section 6, we extend the D^2VD algorithm to a nonlinear version, namely Nonlinear D^2VD algorithm. Section 7 gives the experimental results on both synthetic and real-world datasets. Finally, Section 8 concludes the paper.

2 Related Work

To make a causal inference, the golden standard approaches are randomized experiments, such as A/B testing [3], where different treatments are randomly assigned to units. However, fully randomized experiments are frequently infeasible [4], [5] on many occasions due to high costs or ethical reasons.

Rosenbaum and Rubin [6] proposed a causal inference framework for causal inference in observational studies, based on the propensity score which is estimated via logistic regression. Then, many other machine learning algorithms (e.g. gradient boosted machine, bagged CART and random forest) are employed to estimate the propensity score [22], [23]. And various methods have been proposed based on the propensity score, including the propensity score matching, stratification on the propensity score, inverse propensity



(a) Neyman-Rubin's Causal Framework. (b) Our Causal Framework.

Fig. 2. Comparison between Neyman-Rubin's framework and our proposed causal framework. (a) Neyman-Rubin's framework treats all observed variables $\mathbf{U}$ as confounders $\mathbf{X}$. (b) In our causal framework, we separate the variables $\mathbf{U}$ into three different partitions: (1) Confounders $\mathbf{X}$, which are associated with the treatment T and may be causally related to the outcome Y, (2) Adjustment Variables $\mathbf{Z}$, which are causally related to outcome Y, but independent with treatment T, and (3) Irrelevant Variables $\mathbf{I}$ (Omitted), which are independent with both treatment T and outcome Y.

weighting and regression on propensity score [6], [7], [8], [24]. The inverse propensity weighting is a commonly used method to control confounders for causal inference in observational studies. It was first proposed in [25] and had been part of a large family of causal models known as the marginal structural model [10], [11]. With a combination between inverse propensity weighting and regression, Bang and Robins [12] proposed a doubly robust estimator, which is proven to have a smaller asymptotic variance [26]. The inverse propensity weighting and doubly robust methods have been widely used in various fields, including economics [13], epidemiology [14], health care [27], social science [16], political science [26] and advertising [24].

Our work is distinct from these existing works which assume all variables as confounders and adopt the causal framework shown in Fig. 2a. We propose to separate the confounders and adjustment variables and redesign a new causal framework shown in Fig. 2b, utilizing the adjustment variables to reduce the variance of estimated ATE. Furthermore, we propose a data-driven approach to automatically separate confounders and adjustment variables and simultaneously estimate treatment effect in observational studies.

Our work is closely related to [21], which reduced the variance of estimated ATE in randomized experiments by adjusting covariates with lasso regularizers. But our work differs from [21] in that [21] is tailored for randomized experiments and there is no need to deal with confounding in randomized experiments. In observational studies, we need to control for confounding and try to reduce variance by including adjustment variables in the meantime. The crux of our paper is to wisely separate the two sets of variables automatically.

Comparing to the preliminary version [28], this one comprises a substantial amount of additional theoretical, algorithmic and experimental efforts and contributions. Key points of differences lie in the following aspects: First, as the proposed estimator in our conference paper, we give theoretical analysis on both bias and variance of the estimated treatment effect. Second, to address the challenges from high dimensional variables and non-linear regression, we extend our D^{2}VD algorithm to a Non-linear version, namely a Non-linear D^{2}VD algorithm for treatment effect estimation. Third, we report a series of statistical tests that examine the performance of the Non-linear D^{2}VD algorithm for treatment effect estimation and find that the method achieves more precise results than D^{2}VD algorithm, especially in the settings with nonlinear regression.

3 Adjusted ATE Estimator

In this section, we first give the notations and assumptions for the ATE estimation in observational studies, then propose a new adjusted ATE estimator by utilizing the adjustment variables for reducing the variance of estimated ATE.

3.1 Notations and Assumptions

As described in our causal diagram in Fig. 2b, we define a treatment as a random variable T and a potential outcome as $Y(t)$ which corresponds to a specific treatment $T = t$. In this paper, we only consider binary treatment, that is $t \in \{0, 1\}$. We define the units which received the treatment, which is $T = 1$, as treated units and the others with $T = 0$ as control units. Then for each unit indexed by $i = 1, 2, \ldots, n$, we observe a treatment T_i, an outcome Y_i^{obs} and a vector of variables U_i. Our observed outcome Y_i^{obs} of unit i can be denoted by:

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

We use S_a and S_b to represent the units set with treated status ($T = 1$) and control status ($T = 0$), respectively.

In our paper, for any column vector $\mathbf{v} = (v_1, v_2, \ldots, v_p)^T$, let $\|\mathbf{v}\|_2^2 = \sum_{i=1}^p v_i^2$, and $\|\mathbf{v}\|_1 = \sum_{i=1}^p |v_i|$. Let $\odot$ and $\langle \cdot, \cdot \rangle$ refer to the Hadamard product and dot product of two vectors, respectively. More symbols are summarized in Table 1.

In observational studies, there are three standard assumptions [6] for ATE estimation.

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

Assumption 2 (Unconfoundedness). *The distribution of treatment T_i for each unit i is independent of its potential*

TABLE 1
Symbols and Definitions

Symbol	Definition
n	sample size
p	dimension of variables
$\mathbf{X} \in \mathbb{R}^{n \times p}$	confounders
$\mathbf{Z} \in \mathbb{R}^{n \times p}$	adjustment variables
$\mathbf{U} \in \mathbb{R}^{n \times p}$	all variables, including $\mathbf{X}$ and $\mathbf{Z}$
T	treatment
Y	potential outcome
Y^{obs}	observed outcome
$Y^{\star}$	transformed outcome
Y^{+}	adjusted transformed outcome
S_a	units set with treated status ($T = 1$)
S_b	units set with control status ($T = 0$)

outcome $(Y_i(0), Y_i(1))$, *given its observed variables* $\mathbf{U}_i$. *Formally,* $T_i \perp (Y_i(0), Y_i(1)) \,|\, \mathbf{U}_i$ *or* $p((Y_i(0), Y_i(1)) \,|\, T_i, \mathbf{U}_i) = p((Y_i(0), Y_i(1)) \,|\, \mathbf{U}_i)$.[2]

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

3.2 Adjusted ATE Estimator

The important goal of causal inference in observational studies is to evaluate the ATE on outcome Y. The ATE represents the mean (average) difference between the potential outcome of units under treated and control status. Formally, the ATE is defined as:

$$ATE = E\big[Y(T=1) - Y(T=0)\big], \tag{2}$$

where $E(\cdot)$ refers to the expectation function.

The Eq. (2) is infeasible, because of "the counterfactual problem" [24]. That is for each unit, we can only observe one potential outcome corresponding to its treatment status, treated or control.

One can address this counterfactual problem by approximating the unobserved potential outcome. The simplest approach is to directly compare the average outcome between the treated and control units. In observational studies, however, comparing two groups of units directly is likely to have a bias if the treatment assignment is not random, as confounding impact is not taken into account [24].

To unbiasedly evaluate the ATE in observational studies, one has to control the impact of confounders. Under the assumptions (1,2,3), [6] introduced the propensity score to summarize the information required to control the confounders. The propensity score, denoted by $e(\mathbf{U})$, was defined as the probability of units to be in treated status, i.e., $T = 1$, when given all variables $\mathbf{U}$. Actually, from our causal framework in Fig. 2b, we know that only confounders $\mathbf{X}$ are associated with the treatment, therefore

$$e(\mathbf{U}) = p(T=1|\mathbf{U}) = p(T=1|\mathbf{X}) = e(\mathbf{X}). \tag{3}$$

Based on the propensity score, [25] proposed the transformed outcome $Y^{\star}$ to address the counterfactual problem in Eq. (2) with Inverse Propensity Weighting (IPW) estimator $\widehat{ATE}_{IPW}$, see also [29]. Specifically, the treated units $(T = 1)$ is weighted by $\frac{T}{e(U)}$ and estimate $\hat{E}(Y(T=1)) = \frac{T \cdot Y^{obs}}{e(U)}$, the control units is weighted by $\frac{1-T}{1-e(U)}$ and estimate $\hat{E}(Y(T=0)) = \frac{(1-T)\cdot Y^{obs}}{1-e(U)}$. Then, the transformed outcome $Y^{\star}$ is defined as

$$\begin{aligned} Y^{\star} &= \hat{E}(Y(T=1)) - \hat{E}(Y(T=0)) \\ &= \frac{T \cdot Y^{obs}}{e(\mathbf{U})} - \frac{(1-T)\cdot Y^{obs}}{1-e(\mathbf{U})} \\ &= Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} \\ &= \end{aligned} \tag{4}$$

and the IPW estimator is defined as

$$\widehat{ATE}_{IPW} = \widehat{E}(Y^{\star}) = \widehat{E}\left(Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right). \tag{5}$$

However, most previous approaches based on propensity score usually treat all observed variables as confounders when estimating the propensity score. This will make the estimated treatment effect imprecise and with inflated variance because some variables could be non-confounders and have a direct impact on the outcome.

Therefore, based on our causal diagram as shown in Fig. 2b, we propose to separate all observed variables $\mathbf{U}$ into three sets, the confounders $\mathbf{X}$, the adjustment variables $\mathbf{Z}$ and irrelevant variables $\mathbf{I}$ (Omitted in Fig. 2b). And then, we propose a newly adjusted estimator by incorporating adjustment variables to reduce the variance of estimated ATE under the following assumption.

Assumption 4 (Separateness). *The observed variables* $\mathbf{U}$ *can be decomposed into three sets, that is* $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I})$, *where confounders* $\mathbf{X}$ *are associated with the treatment* T *and might be correlated to the outcome* Y; *adjustment variables* $\mathbf{Z}$ *are causally related to outcome* Y *but independent with treatment* T; *and irrelevant variables* $\mathbf{I}$ *are independent with both treatment* T *and outcome* Y.

With assumption 4, we define our adjusted transformed outcome Y^{+} based on $Y^{\star}$ as

$$Y^{+} = \big(Y^{obs} - \boldsymbol{\phi}(\mathbf{Z})\big) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}, \tag{6}$$

where $\boldsymbol{\phi}(\mathbf{Z})$ is associated with $\mathbf{Z}$ and refers to the effect of $\mathbf{Z}$ on Y. $(Y^{obs} - \boldsymbol{\phi}(\mathbf{Z}))$ helps to reduce the variance among Y.

Then we propose the adjusted estimator $\widehat{ATE}_{adj}$ as

$$\widehat{ATE}_{adj} = \widehat{E}(Y^{+}) = \widehat{E}\left(\big(Y^{obs} - \boldsymbol{\phi}(\mathbf{Z})\big) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right). \tag{7}$$

In the next section, we will prove that our proposed adjusted estimator can unbiasedly estimate treatment effect and achieve more robust results than those estimators which assume all variables as confounders.

4 Theoretical Analysis

In this section, we give the theoretical analysis about our adjusted estimator $\widehat{ATE}_{adj}$, including bias analysis and variance analysis.

4.1 Bias Analysis

For our adjusted transformed outcome Y^{+} in Eq. (6), we have following property.

Theorem 1. *Under assumptions 1-4, we have*

$$E(Y^{+}|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z}). \tag{8}$$

Proof. First, under Assumption 4,

$$E\left(\boldsymbol{\phi}(\mathbf{Z}) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} |\mathbf{X}, \mathbf{Z}\right) \tag{9}$$

2. It means all the covariates/confounders that are related to both the treatment T and the outcome $(Y(0), Y(1))$ have been included in the observed variables $\mathbf{U}$ for all units.

$$= \frac{\phi(\mathbf{Z})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \cdot E(T - e(\mathbf{X})|\mathbf{X}, \mathbf{Z}) \tag{10}$$

$$= \frac{\phi(\mathbf{Z})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \cdot (E(T|\mathbf{X}) - e(\mathbf{X})) \tag{11}$$

$$= \frac{\phi(\mathbf{Z})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \cdot (e(\mathbf{X}) - e(\mathbf{X})) = 0. \tag{12}$$

From Eqs. (9) and (10): $\phi(\mathbf{Z})$, $e(\mathbf{X})$ and $1 - e(\mathbf{X})$ are constant when given $\mathbf{X}, \mathbf{Z}$. From Eqs. (10) and (11): Under assumption 4, T is independent with $\mathbf{Z}$, hence $E(T|\mathbf{X}, \mathbf{Z}) = E(T|\mathbf{X})$. Then, with the definition of propensity score that $e(\mathbf{X}) = E(T|\mathbf{X})$, we can obtain Eq. (12) from Eq. (11).

Second, it has been shown in the literature that $E(Y^\star|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z})$, see, e.g., [29]. Then we can derive

$$\begin{aligned} &E(Y^+|\mathbf{X}, \mathbf{Z}) \\ &= E(Y^\star|\mathbf{X}, \mathbf{Z}) - E\left(\phi(\mathbf{Z}) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}|\mathbf{X}, \mathbf{Z}\right) \\ &= E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z}). \end{aligned}$$

Then we conclude that $E(Y^+|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z})$. □

With theorem 1, we can obtain following equation

$$E(Y^+) = E(Y(1)) - E(Y(0)). \tag{13}$$

Therefore, our proposed adjusted estimator is unbiased for treatment effect estimation in observational studies.

4.2 Variance Analysis

With our adjusted transformed outcome Y^+ as defined in Eq. (6), we can rewrite the observed outcome Y_i^{obs} for each unit i as

$$Y_i^{obs} = Y_i^+ \cdot \frac{e(\mathbf{X}_i) \cdot (1 - e(\mathbf{X}_i))}{T_i - e(\mathbf{X}_i)} + \phi(\mathbf{Z}_i) + e_i^+, \tag{14}$$

where e_i^+ is the unit-level residuals represented with Y_i^+.

Similarly, we can rewrite the observed outcome Y_i^{obs} with the definition of transformed outcome Y^* in Eq. (4) for each unit i as:

$$Y_i^{obs} = Y^\star \cdot \frac{e(\mathbf{X}_i) \cdot (1 - e(\mathbf{X}_i))}{T - e(\mathbf{X}_i)} + e_i^\star, \tag{15}$$

where $e_i^\star$ is the unit-level residuals represented with $Y_i^\star$.

Condition 1. *There exists a fixed constant $L > 0$ such that for all $n = 1, 2, \ldots$*

$$\frac{1}{n}\sum_{i=1}^{n}(e_i^+)^4 < L, \frac{1}{n}\sum_{i=1}^{n}(e_i^\star)^4 < L,$$

Condition 2. *The means $n^{-1}\sum_{i=1}^{n}(e_i^+)^2$, $n^{-1}\sum_{i=1}^{n}(e_i^\star)^2$, and $n^{-1}\sum_{i=1}^{n} e_i^+ e_i^\star$ converge to finite limits.*

We denote by σ_{adj}^2 the asymptotic variance of our adjusted estimator $\widehat{ATE}_{adj}$, then, under conditions 1 & 2, and according to Bloniarz *et al.* [21], we have

$$\sqrt{|S_a| + |S_b|}\left(\widehat{ATE}_{adj} - ATE\right) \to \mathcal{N}(0, \sigma_{adj}^2),$$

where S_a and S_b represent the units set with treated and control status, respectively.

$$\sigma_{adj}^2 = \lim_{|S_a|+|S_b|\to\infty}\left(\frac{|S_b|}{|S_a|}\sigma_{e^{+(1)}}^2 + \frac{|S_a|}{|S_b|}\sigma_{e^{+(0)}}^2 + 2\sigma_{e^{+(1)}e^{+(0)}}\right),$$

where

$$\begin{aligned} \sigma_{e^{+(1)}}^2 &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_a}\left(e_i^{+(1)}\right)^2, \\ \sigma_{e^{+(0)}}^2 &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_b}\left(e_i^{+(0)}\right)^2, \\ \sigma_{e^{+(1)}e^{+(0)}} &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_a\cup S_b} e_i^{+(1)} e_i^{+(0)}, \end{aligned}$$

and $e_i^{+(1)}$ and $e_i^{+(0)}$ represent the unit-level residuals e_i^+ (see Eq. (14)) of treated and control unit, respectively.

Similarity, denoting by σ_{IPW}^2 as the asymptotic variance of estimators based on Inverse Propensity Weighting (we call them as IPW estimators $\widehat{ATE}_{IPW}$ in this paper), we have

$$\sigma_{IPW}^2 = \lim_{|S_a|+|S_b|\to\infty}\left(\frac{|S_b|}{|S_a|}\sigma_{e^\star(1)}^2 + \frac{|S_a|}{|S_b|}\sigma_{e^\star(0)}^2 + 2\sigma_{e^\star(1)e^\star(0)}\right),$$

where $e_i^{\star(1)}$ and $e_i^{\star(0)}$ represent the unit-level residuals $e_i^\star$ (see Eq. (15)) of treated and control unit, respectively.

With the σ_{adj}^2 and σ_{IPW}^2, we have following theorem.

Theorem 2. *The asymptotic variance of our adjusted estimator $\widehat{ATE}_{adj}$ is no greater than IPW estimator $\widehat{ATE}_{IPW}$:*

$$\sigma_{adj}^2 \le \sigma_{IPW}^2.$$

Proof. The difference between σ_{adj}^2 and σ_{IPW}^2 is

$$\begin{aligned} \sigma_{adj}^2 - \sigma_{IPW}^2 = &\frac{|S_b|}{|S_a|}\lim_{|S_a|+|S_b|\to\infty}\left(\sigma_{e^{+(1)}}^2 - \sigma_{e^\star(1)}^2\right) \\ &+ \frac{|S_a|}{|S_b|}\lim_{|S_a|+|S_b|\to\infty}\left(\sigma_{e^{+(0)}}^2 - \sigma_{e^\star(0)}^2\right) \\ &+ 2\lim_{|S_a|+|S_b|\to\infty}\left(\sigma_{e^{+(1)}e^{+(0)}} - \sigma_{e^\star(1)e^\star(0)}\right). \end{aligned} \tag{16}$$

With the orthogonal regularizer between α and β in Eq. (23), we can obtain $\phi(\mathbf{Z})$ and $e(\mathbf{X})$ are the orthogonal projections of the outcomes, we have

$$\left(Y^\star \cdot \frac{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}{T - e(\mathbf{X})}\right)^{\mathrm{T}} \phi(\mathbf{Z}) = 0,$$

$$\left(Y^+ \cdot \frac{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}{T - e(\mathbf{X})}\right)^{\mathrm{T}} \phi(\mathbf{Z}) = 0,$$

Then we have the following three equalities:

$$\begin{aligned}
\sigma^2_{e^{+(1)}} - \sigma^2_{e^{\star}(1)} &= ||e^{+(1)}||_2^2 - ||e^{\star}(1)||_2^2 = -||\phi(\mathbf{Z})||_2^2 \leq 0,\\
\sigma^2_{e^{+(0)}} - \sigma^2_{e^{\star}(0)} &= ||e^{+(0)}||_2^2 - ||e^{\star}(0)||_2^2 = -||\phi(\mathbf{Z})||_2^2 \leq 0,\\
\sigma_{e^{+(1)}e^{+(0)}} - \sigma_{e^{\star}(1)e^{\star}(0)} &= \left(e^{+(1)}\right)^{\mathrm{T}}\left(e^{+(0)}\right) - \left(e^{\star}(1)\right)^{\mathrm{T}}\left(e^{\star}(0)\right)\\
&= -||\phi(\mathbf{Z})||_2^2 \leq 0.
\end{aligned}$$

And with Equation (16), we conclude that

$$\sigma^2_{adj} - \sigma^2_{IPW} \leq 0.$$

□

Therefore, our adjusted estimator is unbiased and has a smaller variance than those estimators which assume all variables as confounders, such as the IPW estimator. This enables us to construct tighter confidence intervals for the true ATE.

5 D²VD Algorithm and Optimization

In this section, we give the details of our Data-Driven Variable Decomposition (D²VD) algorithm for automatically separating confounders and simultaneously estimating the treatment effect. We also introduce the parameter tuning method for the "no ground truth" problem in observational causal inference.

5.1 D²VD Algorithm

With our adjusted estimator in Eq. (7), we can obtain estimated ATE by regressing our adjusted transformed outcome Y^+ against the variables $\mathbf{U}$ and minimizing the following objective function

$$\min_{\phi(\cdot),e(\cdot),h(\cdot)} \quad \|Y^+ - h(\mathbf{U})\|^2, \tag{17}$$

where $h(\mathbf{U})$ denotes the effect of $\mathbf{U}$ on our adjusted transformed outcome Y^+ with a function h. Then, the estimated ATE by our adjusted estimator $\widehat{ATE}_{adj}$ can be obtained by $E(h(\mathbf{U}))$.

In practice, we specify $\phi(\mathbf{Z})$ and $h(\mathbf{U})$ as linear functions[3] with coefficient vector α and γ, that is

$$\phi(\mathbf{Z}) = \mathbf{Z}\alpha, \tag{18}$$

$$h(\mathbf{U}) = \mathbf{U}\gamma, \tag{19}$$

and adopt the linear-logistic regression to evaluate the propensity score $e(\mathbf{X})$ with coefficient vector β:

$$e(\mathbf{X}) = p(T=1|\mathbf{X}) = \frac{1}{1+\exp(-\mathbf{X}\beta)}. \tag{20}$$

For simplifying the equations and formulations in our paper, we involve the inverse propensity weighting function $W(\beta)$ as a function of the coefficient vector β:

$$\begin{aligned}
W(\beta) &:= \frac{T - e(\mathbf{X})}{e(\mathbf{X})\cdot(1-e(\mathbf{X}))}\\
&= (2T - \mathbf{1}_n) \odot \left(\mathbf{1}_n + \exp\Big((\mathbf{1}_n - 2T)\odot \mathbf{X}\beta\Big)\right)
\end{aligned} \tag{21}$$

where $\odot$ denoted the Hadamard product and $\mathbf{1}_n = [\underbrace{1, 1, \ldots, 1}_{n}]^T$.

In the specifications of Eqs. (18), (20), and (21), we have assumed the knowledge of the decomposition $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I})$. Nevertheless, we don't know the exact separation in practice. Hence we use the full set of candidate variables $\mathbf{U}$ to replace $\mathbf{X}$ and $\mathbf{Z}$ instead, and use automatic algorithm to achieve the separation. We update our objective function in Eq. (17) as:

$$\begin{aligned}
\min \quad & \|(Y^{obs} - \mathbf{U}\alpha)\cdot W(\beta) - \mathbf{U}\gamma\|_2^2,\\
s.t. \quad & \|\alpha\|_1 \leq \lambda,\\
& \|\beta\|_1 \leq \delta,\\
& \|\gamma\|_1 \leq \eta,\\
& \langle\alpha, \beta\rangle = 0.
\end{aligned} \tag{22}$$

where the coefficient vector α is optimized for separating the adjustment variables $\mathbf{Z}$ and β is for separating confounders $\mathbf{X}$ form variables $\mathbf{U}$. We can obtain the ATE by estimating $E(\mathbf{U}\gamma)$. That is, with the optimized coefficient vectors α, β and γ, we can separate the confounders and adjustment variables and simultaneously estimate the ATE.

In particular, we employ an orthogonal regularizer on α and β for ensuring the separation of confounders and adjustment variables. Besides, we add L_1 penalties on α, β and γ to eliminate irrelevant variables $\mathbf{I}$ and to further reduce variance and address the sparseness problem of variables.

Then with the lagrangian from Eq. (22), we can rewrite our objective function, denoted as $\mathcal{J}(\alpha, \beta, \gamma)$, with the penalty regularizers as:

$$\begin{aligned}
\mathcal{J}(\alpha,\beta,\gamma) = &\|(Y^{obs} - \mathbf{U}\alpha)\cdot W(\beta) - \mathbf{U}\gamma\|_2^2\\
&+ \lambda\|\alpha\|_1 + \delta\|\beta\|_1 + \eta\|\gamma\|_1 + \mu\|\alpha^T\beta\|_2^2.
\end{aligned} \tag{23}$$

To minimize the objective function $\mathcal{J}(\alpha, \beta, \gamma)$ with L_1 norm regularizer which is non-smooth and undifferentiable,, we adopt the proximal gradient algorithm [30]. For each optimizing step in proximal gradient algorithm, we used the proximal operator [30] for L_1 norm regularizer.

First, we split our objective function $\mathcal{J}(\alpha, \beta, \gamma)$ into two parts: the differentiable part $f(\alpha, \beta, \gamma)$ and the undifferentiable part $g(\alpha, \beta, \gamma)$ (simplified as $\mathcal{J}$, f and g). That is

$$\mathcal{J} = f(\alpha,\beta,\gamma) + g(\alpha,\beta,\gamma), \tag{24}$$

$$f = \|(Y^{obs} - \mathbf{U}\alpha)\cdot W(\beta) - \mathbf{U}\gamma\|_2^2 + \mu\|\alpha^T\beta\|_2^2, \tag{25}$$

$$g = \lambda\|\alpha\|_1 + \delta\|\beta\|_1 + \eta\|\gamma\|_1. \tag{26}$$

Then with the operator splitting property [30] of proximal gradient algorithm, we can get the optimized parameter (i.e., $\alpha^{(t+1)}$) at the tth iteration by proximal operator $prox_{\kappa^{(t)}g}$ of function $g(\cdot)$ with the step size $\kappa^{(t)}$:

3. In practice, higher order terms or interaction terms of the original variables can be included in $\mathbf{U}$. Thus the linearity assumption is not as stringent as it seems.

$$\alpha^{(t+1)} = prox_{\kappa^{(t)}g}\left(\alpha^{(t)} - \kappa^{(t)}\frac{\partial f(\cdot)}{\partial \alpha}\right), \tag{27}$$

where $\frac{\partial f(\cdot)}{\partial \alpha}$ refers to the gradient of function $f(\cdot)$ on the variable α and we set

$$\begin{aligned} prox_{\kappa^{(t)}g}(x) &= (x - \kappa^{(t)}\cdot\lambda)_+ - (-x - \kappa^{(t)}\cdot\lambda)_+ \\ &= \begin{cases} x_i - \kappa^{(t)}\cdot\lambda & x_i \geq \kappa^{(t)}\cdot\lambda \\ 0 & |x_i| \leq \kappa^{(t)}\cdot\lambda \\ x_i + \kappa\cdot\lambda & x_i \leq -\kappa^{(t)}\cdot\lambda \end{cases} \end{aligned} \tag{28}$$

The λ in Eq. (28) is the coefficient of variable α in function $g(\cdot)$. If the optimized parameter is β, then it should be δ and should be η for optimized parameter γ.

With the proximal gradient algorithm, we can minimize the objective function in Eq. (23). That is, starting from some random initialization on α, β, γ, we solve each of them alternatively with the other two parameters as fixed and step by step until convergence. Obviously, the objective function $\mathcal{J}(\alpha, \beta, \gamma)$ is bounded below by 0 and the alternating proximal gradient search procedure will reduce it monotonically, the algorithm is guaranteed to be convergent. Specifically, the gradients of the function $f(\alpha, \beta, \gamma)$ with the respect to the variables are:

$$\frac{\partial f(\cdot)}{\partial \alpha} = -2(W(\beta)\cdot\mathbf{1}_p^{\mathrm{T}} \odot \mathbf{U})^{\mathrm{T}}$$

$$\frac{\partial f(\cdot)}{\partial \beta} = 2\left((Y - \mathbf{U}\alpha)\cdot\mathbf{1}_p^{\mathrm{T}} \odot \frac{\partial W(\beta)}{\partial \beta}\right)^{\mathrm{T}}$$

$$\frac{\partial f(\cdot)}{\partial \gamma} = -2\mathbf{U}^{\mathrm{T}}\cdot\left((Y - \mathbf{U}\alpha)\odot W(\beta) - \mathbf{U}\gamma\right).$$

where

$$\begin{aligned} \frac{\partial W(\beta)}{\partial \beta} &= (2T - \mathbf{1}_n)\odot\exp\left((\mathbf{1}_n - 2T)\odot\mathbf{U}\beta\right) \\ &\quad \odot(\mathbf{1}_n - 2T)\cdot\mathbf{1}_p^T \odot \mathbf{U}, \end{aligned}$$

and $\mathbf{1}_n = [\underbrace{1, 1, \ldots, 1}_{n}]^{\mathrm{T}}$, $\mathbf{1}_p = [\underbrace{1, 1, \ldots, 1}_{p}]^{\mathrm{T}}$.

Thus, we apply the following proximal gradient based approach on our Data-Driven Variable Decomposition algorithm as describe in Algorithm 1.

The function $\hat{f}_\kappa(\cdot)$ in Algorithm 1 is defined as:

$$\hat{f}_\kappa(x, y) = f(y) + (x - y) + \frac{\partial f(\cdot)}{\partial x}^{\mathrm{T}}(1/(2\kappa))\|x - y\|_2^2. \tag{29}$$

Our model can be applied in the real system to deal with the causal inference problem in observational studies.

5.2 Complexity Analysis

By analyzing the procedure of optimization in Algorithm 1, we know the main cost is to calculate the loss $\mathcal{J}(\alpha, \beta, \gamma)$ and update those parameters including α, β, and γ. We analyze the time complexity of each of them respectively. For the calculation of the loss, its complexity is $O(np)$, where n is the sample size and p is the dimension of observed variables. The complexity of updating parameter α is dominated by the step of calculating the partial gradients of function $f(\cdot)$ with respect to variable α. The complexity of $\frac{\partial f(\cdot)}{\partial \alpha}$ is $O(np)$. For updating parameter β, it complexity is dominated by the step of calculating $\frac{\partial f(\cdot)}{\partial \beta}$ and $\frac{\partial W(\beta)}{\partial \alpha}$, their complexity are also $O(np)$. Similarly, we can find the complexity of updating parameter γ is also $O(np)$.

In total, the complexity of each iteration in Algorithm 1 is $O(np)$.

5.3 Parameters Tuning

The main challenge of parameter tuning for causal inference algorithms in observational studies is that there is no ground truth about the true ATE.

Algorithm 1. Data-Driven Variable Decomposition Algorithm

Require: Initialization $\mathcal{J}^{(0)} = \mathcal{J}(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})$.
Ensure: $\mathcal{J}^{(0)} \geq 0, \mathcal{J}^{(t+1)} < \mathcal{J}^{(t)}$
  for $t = 1, 2, \cdots$ **do**
    Calculate $\frac{\partial f(\cdot)}{\partial \alpha}$, $\frac{\partial f(\cdot)}{\partial \beta}$ and $\frac{\partial f(\cdot)}{\partial \gamma}$
    $\kappa = 1$
    while 1 **do**
      Let $\alpha^{(t+1)} = prox_{\kappa g}\left(\alpha^{(t)} - \kappa\frac{\partial f(\cdot)}{\partial \alpha}\right)$
      break if $f(\alpha^{(t+1)}) \leq \hat{f}_\kappa(\alpha^{(t+1)}, \alpha^{(t)})$
      Update $\kappa = \frac{1}{2}\kappa$
    end while
    $\kappa = 1$
    while 1 **do**
      Let $\beta^{(t+1)} = prox_{\kappa g}\left(\beta^{(t)} - \kappa\frac{\partial f(\cdot)}{\partial \beta}\right)$
      break if $f(\beta^{(t+1)}) \leq \hat{f}_\kappa(\beta^{(t+1)}, \beta^{(t)})$
      Update $\kappa = \frac{1}{2}\kappa$
    end while
    $\kappa = 1$
    while 1 **do**
      Let $\gamma^{(t+1)} = prox_{\kappa g}\left(\gamma^{(t)} - \kappa\frac{\partial f(\cdot)}{\partial \gamma}\right)$
      break if $f(\gamma^{(t+1)}) \leq \hat{f}_\kappa(\gamma^{(t+1)}, \gamma^{(t)})$
      Update $\kappa = \frac{1}{2}\kappa$
    end while

$$\mathcal{J}^{(t+1)} = \mathcal{J}(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})$$

  end for

To address this challenge, we employed the matching method to evaluate the ATE and set it as "approximal ground truth" like Athey and Imbens did in [29]. Specifically, for each unit i, find its closest match among the units with opposite treatment status in the test dataset:

$$match(i) = \arg\min_{j:T_j = 1 - T_i} \|U_i - U_j\|_2^2. \tag{30}$$

We can estimate the true ATE with matching estimator $ATE_{matching}$ by comparing the average outcome between the matched treated and control units sets. We set it as "approximal ground truth". With the "approximal ground truth", we can tune parameters of our algorithm with cross-validation by grid searching.

6 Non-Linear D^{2}VD Algorithm

In previous section, we assume the function between adjustment variables $\mathbf{Z}$ and outcome Y, and function between confounders $\mathbf{X}$ and treatment T are linear and linear-logistic in Eqs. (18) and (20). In practice, those functions might be non-linear structures in real applications.

To address these challenges, we adopt a neural network to learn a non-linear representation of variables and finally extend our D^{2}VD algorithm to a non-linear version, namely Non-linear D^{2}VD Algorithm. To make sure we can correctly specify the propensity score model, which is very important for propensity score based methods, we keep the function between confounders $\mathbf{X}$ and treatment T as linear-logistic, but the function between adjustment variables $\mathbf{Z}$ and outcome Y is non-linear. Then, the objective function of our N-D^{2}VD algorithm can be written as:

$$\begin{aligned}\mathcal{J}(\mathbf{A},\alpha,\beta,\gamma) = &\|(Y^{obs} - h(\mathbf{U},\mathbf{A})\alpha)\cdot W(\beta) - \mathbf{U}\gamma\|_2^2 \\ &+ \lambda\|\alpha\|_1 + \delta\|\beta\|_1 + \eta\|\gamma\|_1 + \mu\|\beta^T\mathbf{A}^{(1)}\|_1.\end{aligned} \tag{31}$$

where $\mathbf{A}$ refers to the parameters in neural network for learning non-linear representation of variables $h(\mathbf{U},\mathbf{A})$. Given the input $\mathbf{U}$, the hidden representations for each layer in neural network are shown as follows:

$$\begin{aligned}h(\mathbf{U})^{(1)} &= \sigma(\mathbf{U}\mathbf{A}^{(1)} + b^{(1)}) \\ h(\mathbf{U})^{(k)} &= \sigma(h(\mathbf{U})^{(k-1)}\mathbf{A}^{(k)} + b^{(k)}), k = 2,\ldots,K\end{aligned} \tag{32}$$

where K is the number of layer. $\mathbf{A}^{(k)}$ and $b^{(k)}$ are weight matrix and bias on kth layer. $\sigma(\cdot)$ represents non-linear activation function[4]. $h(\mathbf{U},\mathbf{A}) = h(\mathbf{U},\mathbf{A})^{(K)}$ is the hidden representation of final layer.

In Eq. (31), the orthogonal constraint $\|\beta^T\mathbf{A}^{(1)}\|_1$ helps to make sure the separation of confounders and adjustment variables, where $\mathbf{A}^{(1)}$ is the weight matrix in the first layer of neural network for learning non-linear representation of adjustment variables as we have shown in Eq. (32), and β is the coefficient of confounders for estimating propensity score as we have shown in Eq. (20).

Finally, with the optimized parameters $\mathbf{A}^{(1)}$, β, and γ, we can separate the confounders and adjustment variables, and simultaneously estimate the ATE. Note that, we only consider the non-linear function between adjustment variables $\mathbf{Z}$ and outcome Y in our N-D^{2}VD algorithm, the non-linear function between confounders $\mathbf{X}$ and treatment T can also be easily incorporated.

7 Experiments

In this section, we check the performance of our proposed algorithms on treatment effect estimation with multiple synthetic datasets and two real-world datasets.

7.1 Baseline Estimators

We implement the following baseline estimators to evaluate the ATE for comparison.

- *Direct Estimator* $\widehat{ATE}_{DIR}$: It evaluates the ATE by directly comparing the average outcome between the treated and control units. It ignores the confounding effect of confounders on treatment.
- *Linear Regression with LASSO* $\widehat{ATE}_{LASSO}$[31]: It directly regresses outcome on treatment and observed variables, and evaluates the ATE with the regression coefficient of treatment. It ignores the confounding effect between confounders and treatment.
- *IPW Estimator* $\widehat{ATE}_{IPW}$ [6]: It evaluates the ATE via reweighting observations with inverse of propensity score. It treats all variables as confounders and ignores the adjustment variables.
- *Doubly Robust Estimator* $\widehat{ATE}_{DR}$ [12]: It evaluates the ATE by combination of IPW and regression methods. It ignores the separation of confounders and adjustment variables.

In this paper, we implemented $\widehat{ATE}_{IPW}$ and $\widehat{ATE}_{DR}$ with *lasso regression* [31] for variables selection in high dimensional settings.

7.2 Evaluation Metrics

In our experiments, we perform the task of treatment effect estimation. To evaluate the performance of our proposed methods, we carry out the experiments for 50 times independently. Based on the estimated treatment effect ($\widehat{ATE}$) in each experiment, we calculate and report its $Bias$, standard deviations (SD), mean absolute errors (MAE) and root mean square errors ($RMSE$) with following definitions:

$$\begin{aligned}Bias &= |\frac{1}{K}\sum_{k=1}^{K}\widehat{ATE}_k - ATE| \\ SD &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATE}_k - \frac{1}{K}\sum_{k=1}^{K}\widehat{ATE}_k)^2} \\ MAE &= \frac{1}{K}\sum_{k=1}^{K}|\widehat{ATE}_k - ATE| \\ RMSE &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATE}_k - ATE)^2},\end{aligned}$$

where K is the number of experiments, $\widehat{ATE}_k$ is the estimated ATE in kth experiment and ATE represents the *true treatment effect*.

7.3 Experiments on Synthetic Data

7.3.1 Dataset

To generate the synthetic dataset, we set the sample size $n = \{10000, 20000\}$ and the dimension of observed variables $p = \{50, 100, 150\}$. We first generate the variables $\mathbf{U} = (\mathbf{X},\mathbf{Z},\mathbf{I}) = (\mathbf{x}_1,\ldots,\mathbf{x}_{p_x},\mathbf{z}_1,\ldots,\mathbf{z}_{p_z},\mathbf{i}_1,\ldots,\mathbf{i}_{p_i})$ with independent gaussian distributions as

$$\mathbf{x}_1,\ldots,\mathbf{x}_{p_x},\mathbf{z}_1,\ldots,\mathbf{z}_{p_z},\mathbf{i}_1,\ldots,\mathbf{i}_{p_i} \quad \overset{iid}{\sim} \quad \mathcal{N}(0,1),$$

where p_x, p_z and p_i represent the dimension of confounders $\mathbf{X}$, adjustment variables $\mathbf{Z}$ and irrelevant variables $\mathbf{I}$, respectively. And $p_x = 0.2 * p$, $p_z = 0.2 * p$, $p_i = 0.6 * p$.

4. We use sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ as non-linear activation function.

To test the robustness of all estimators, we generate the binary treatment variable T from a logistic function (T_{logit}) and a misspecified function (T_{missp}) as

$$T_{logit} \sim Bernoulli(1/(1+\exp(-\sum_{i=1}^{p_x} x_i \cdot r))) \; and$$

$$T_{missp} = \begin{cases} 1 & if \;\; \sum_{i=1}^{p_x} x_i/\sqrt{p/5} + \mathcal{N}(0,1) \; > \; 0.5 \\ 0 & otherwise, \end{cases}$$

where $r = 50/p$.

The outcome Y is generated as

$$Y = \sum_{i=1}^{p_x} \mathbf{x}_i \cdot \omega_i + \sum_{j=1}^{p_z} \mathbf{z}_j \mathbf{z}_{j+1} \cdot \rho_j + T + \mathcal{N}(0,1),$$

In synthetic dataset, the features $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p_x})$ are correlated to the treatment and outcome, simulating a confounding effect. From the generation of outcome, we know the *true treatment effect* in synthetic dataset is **1**.

7.3.2 Treatment Effect Estimation

To evaluate the performance of our proposed method, we carry out the experiments 50 times independently. Based on our estimated ATE, we calculate the $Bias$, SD, MAE, and RMSE, and report the results in Table 2, where the smaller $Bias$, SD, MAE and RMSE are better. From Table 2, we have following observations.

- The direct estimator ($\widehat{ATE}_{DIR}$) is failed (with large $Bias$) under all settings because it ignores the confounders, hence did not consider the confounding effect.
- Linear regression ($\widehat{ATE}_{LASSO}$) can improve the accuracy of the treatment effect than the direct method. But its performance is worse than our propose method since it ignores the confounding bias between treatment and observed variables.
- The IPW estimator ($\widehat{ATE}_{IPW}$) can estimate the ATE more precisely (with smaller $Bias$) when $T = T_{logit}$, but with a bigger $Bias$ when propensity score model is misspecified by setting $T = T_{missp}$.
- With a combination of IPW and regression models, the DR estimator ($\widehat{ATE}_{DR}$) can get better performance than the IPW estimator, especially when $T = T_{missp}$, but with higher variance (or SD) than regression model ($\widehat{ATE}_{LASSO}$).
- With considering the separation between confounders and adjustment variables, our D^{2}VD estimator ($\widehat{ATE}_{D^2VD}$) can improve the accuracy (smaller $Bias$) and reduce the variance (smaller SD) for ATE estimation from DR and IPW baseline estimators under different settings.
- By utilizing the neural network to capture the nonlinear structure among variables, our proposed Nonlinear D^{2}VD estimator ($\widehat{ATE}_{N-D^2VD}$) is better than D^{2}VD algorithm, and almost achieves the best performance for treatment effect estimation over all settings.

To deeply demonstrate the advantages of our proposed method, we also show the results of ATE by varying the proportion of confounders and adjustment variables ($\frac{p_x}{p_x+p_z}$)

TABLE 2
Results on Synthetic Dataset in Different Settings

	p	p=50				p=100				p=150			
T/n	Estimator	Bias	SD	MAE	RMSE	Bias	SD	MAE	RMSE	Bias	SD	MAE	RMSE
$T = T_{logit}$ $n = 10000$	$\widehat{ATE}_{DIR}$	1.345	0.201	1.345	1.36	1.744	0.222	1.744	1.758	1.922	0.301	1.922	1.946
	$\widehat{ATE}_{LASSO}$	0.053	0.260	0.206	0.266	0.091	0.362	0.302	0.374	0.100	0.328	0.258	0.343
	$\widehat{ATE}_{IPW}$	0.117	0.867	0.652	0.875	0.031	0.631	0.422	0.632	0.081	0.569	0.491	0.575
	$\widehat{ATE}_{DR}$	0.127	0.839	0.641	0.848	0.028	0.529	0.386	0.53	0.09	0.561	0.49	0.568
	$\widehat{ATE}_{D^2VD}$	0.102	0.283	0.243	0.299	0.112	0.319	0.253	0.336	**0.016**	0.343	0.252	0.34
	$\widehat{ATE}_{N-D^2VD}$	**0.016**	**0.17**	**0.136**	**0.17**	**0.018**	**0.148**	**0.116**	**0.149**	0.027	**0.203**	**0.163**	**0.205**
$T = T_{logit}$ $n = 20000$	$\widehat{ATE}_{DIR}$	1.345	**0.11**	1.345	1.35	1.714	0.171	1.714	1.722	1.858	0.223	1.858	1.872
	$\widehat{ATE}_{LASSO}$	**0.049**	0.185	0.141	0.191	**0.010**	0.219	0.172	0.219	**0.010**	0.227	0.186	0.227
	$\widehat{ATE}_{IPW}$	0.071	0.724	0.476	0.727	0.047	0.467	0.387	0.47	0.034	0.408	0.325	0.409
	$\widehat{ATE}_{DR}$	0.071	0.689	0.472	0.692	0.051	0.439	0.376	0.442	0.026	0.395	0.3	0.395
	$\widehat{ATE}_{D^2VD}$	0.145	0.169	0.18	0.221	0.099	0.208	0.183	0.228	0.177	0.284	0.269	0.332
	$\widehat{ATE}_{N-D^2VD}$	0.054	0.120	**0.109**	**0.131**	0.031	**0.124**	**0.108**	**0.128**	0.017	**0.114**	**0.09**	**0.115**
$T = T_{missp}$ $n = 10000$	$\widehat{ATE}_{DIR}$	1.118	0.196	1.118	1.135	1.52	0.348	1.52	1.559	1.896	0.326	1.896	1.924
	$\widehat{ATE}_{LASSO}$	**0.010**	0.203	0.161	0.203	**0.041**	0.336	0.270	0.338	**0.011**	0.393	0.310	0.393
	$\widehat{ATE}_{IPW}$	0.123	0.301	0.263	0.325	0.177	0.468	0.39	0.5	0.157	0.476	0.394	0.501
	$\widehat{ATE}_{DR}$	0.092	0.308	0.257	0.322	0.054	0.501	0.403	0.504	0.016	0.512	0.421	0.513
	$\widehat{ATE}_{D^2VD}$	0.262	0.158	0.264	0.305	0.186	0.312	0.295	0.361	0.27	0.343	0.347	0.434
	$\widehat{ATE}_{N-D^2VD}$	0.099	**0.075**	**0.106**	**0.125**	0.075	**0.118**	**0.119**	**0.14**	0.084	**0.203**	**0.176**	**0.22**
$T = T_{missp}$ $n = 20000$	$\widehat{ATE}_{DIR}$	1.099	0.135	1.099	1.107	1.525	0.203	1.525	1.539	1.926	0.235	1.926	1.94
	$\widehat{ATE}_{LASSO}$	0.043	0.164	0.144	0.170	0.072	0.257	0.227	0.267	**0.047**	0.254	0.204	0.259
	$\widehat{ATE}_{IPW}$	0.084	0.22	0.182	0.235	0.174	0.26	0.252	0.313	0.269	0.346	0.36	0.438
	$\widehat{ATE}_{DR}$	**0.01**	0.226	0.182	0.226	0.034	0.274	0.232	0.276	0.141	0.355	0.313	0.382
	$\widehat{ATE}_{D^2VD}$	0.275	0.128	0.276	0.303	0.247	0.226	0.285	0.333	0.229	0.253	0.284	0.339
	$\widehat{ATE}_{N-D^2VD}$	0.087	**0.060**	**0.092**	**0.106**	**0.073**	**0.087**	**0.093**	**0.113**	0.060	**0.101**	**0.093**	**0.118**

The Bias refers to the absolute error between the true and estimated ATE. The SD, MAE and RMSE represent the standard deviations, mean absolute errors and root mean square errors of estimated ATE ($\widehat{ATE}$) after 50 times independently experiments, respectively. The smaller Bias, SD, MAE and RMSE, the better.

TABLE 3
Results by Varying the Proportion of Confounders and Adjustment Variables ($\frac{p_x}{p_x+p_z}$) Under Setting $T = T_{logit}, n = 10000, p = 50$, Where $p_x + p_z = 20$

$\mathbf{T} = \mathbf{T_{logit}}, n = 10000, p = 50, p_x + p_z = 20$							
$\frac{p_x}{p_x+p_z}$	Methods	$\widehat{ATE}_{DIR}$	$\widehat{ATE}_{LASSO}$	$\widehat{ATE}_{IPW}$	$\widehat{ATE}_{DR}$	$\widehat{ATE}_{D^2VD}$	$\widehat{ATE}_{N-D^2VD}$
90%	Bias	1.633	0.038	0.062	0.030	**0.014**	0.036
	SD	0.090	0.092	0.286	0.221	0.108	**0.063**
	MAE	1.633	0.080	0.214	0.172	0.089	**0.061**
	RMSE	1.636	0.100	0.293	0.224	0.108	**0.073**
70%	Bias	1.478	0.059	0.093	0.101	0.075	**0.010**
	SD	0.129	0.187	0.346	0.311	0.168	**0.112**
	MAE	1.478	0.165	0.288	0.257	0.147	**0.079**
	RMSE	1.484	0.196	0.358	0.328	0.182	**0.112**
50%	Bias	1.345	0.053	0.117	0.127	0.102	**0.016**
	SD	0.201	0.260	0.867	0.839	0.283	**0.170**
	MAE	1.345	0.206	0.652	0.641	0.243	**0.136**
	RMSE	1.360	0.266	0.875	0.848	0.299	**0.170**
30%	Bias	0.987	0.095	**0.005**	0.016	0.112	0.008
	SD	0.219	0.333	0.630	0.617	0.292	**0.133**
	MAE	0.987	0.281	0.479	0.463	0.256	**0.100**
	RMSE	1.011	0.347	0.630	0.618	0.310	**0.133**
10%	Bias	0.467	0.039	0.081	0.083	0.124	**0.010**
	SD	0.239	0.315	0.356	0.355	0.241	**0.104**
	MAE	0.477	0.256	0.321	0.318	0.217	**0.081**
	RMSE	0.524	0.318	0.365	0.364	0.269	**0.104**

The smaller Bias, SD, MAE and RMSE, the better.

in Table 3. From the results, we find that the bias of $\widehat{ATE}_{DIR}$ increased as increasing in the proportion of confounders $\frac{p_x}{p_x+p_z}$ from 10 to 90 percent, this is because more confounders bring larger confounding bias in data. By comparing with all baselines, our method $\widehat{ATE}_{N-D^2VD}$ achieved the best performance for treatment effect estimation over all settings with different proportions of confounders and adjustment variables.

7.3.3 Variables Decomposition

As we described before, with the optimized $\hat{\alpha}$ and $\hat{\beta}$, our D^{2}VD algorithm can separate the confounders as $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$ and adjustment variables as $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$. To demonstrate the performance of automated variables decomposition of our algorithm, we carry out the experiments 50 times independently and report the true positive rate (TPR) and true negative rate (TNR) in Tables 4 and 5. The formulations of TPR and TNR for separated confounders $\mathbf{X}$ are defined as

$$\text{TPR} = \frac{\#\{\hat{\beta}_i \neq 0, \beta_i \neq 0\}}{\#\{\beta_i \neq 0\}}, \text{TNR} = \frac{\#\{\hat{\beta}_i = 0, \beta_i = 0\}}{\#\{\beta_i = 0\}}. \tag{33}$$

In the same way, we calculate the TPR and TNR for separated adjustment variables $\mathbf{Z}$ via Eq. (33) by using parameter $\hat{\alpha}$.

Similarly, we can obtain the results of variables decomposition of our N-D^{2}VD algorithm with its optimized parameters $\hat{\mathbf{A}}^{(1)}$ and $\hat{\beta}$.

Tables 4 and 5 demonstrate the results of variables decomposition of our D^{2}VD and N-D^{2}VD algorithms, respectively. From Table 4, we know that our D^{2}VD can separate the confounders $\mathbf{X}$ more precisely when $T = T_{logit}$, comparing with $T = T_{missp}$. This is because of the logistic assumption of treatment assignment in our algorithm is correct. Even in the setting $T = T_{missp}$, our D^{2}VD algorithm can still precisely separate the confounders. This enables us to

TABLE 4
Separation Results of Confounders $\mathbf{X}$ and Adjustment Variables $\mathbf{Z}$ From Our D^{2}VD Estimator

$\mathbf{T} = \mathbf{T_{logit}}$							
		$p = 50$		$p = 100$		$p = 150$	
n		TPR	TNR	TPR	TNR	TPR	TNR
$n = 10000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	0.907	1.000
	$\mathbf{Z}$	0.466	0.519	0.536	0.415	0.583	0.363
$n = 20000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	0.970	1.000
	$\mathbf{Z}$	0.296	0.6870	0.389	0.572	0.470	0.504
$\mathbf{T} = \mathbf{T_{missp}}$							
	n	$p = 50$		$p = 100$		$p = 150$	
n		TPR	TNR	TPR	TNR	TPR	TNR
$n = 10000$	$\mathbf{X}$	1.000	1.000	0.915	1.000	0.590	1.000
	$\mathbf{Z}$	0.418	0.545	0.522	0.433	0.587	0.370
$n = 20000$	$\mathbf{X}$	1.000	1.000	0.962	1.000	0.365	1.000
	$\mathbf{Z}$	0.226	0.700	0.388	0.593	0.447	0.497

The closer to **1** *for TPR and TNR is better.*

TABLE 5
Separation Results of Confounders $\mathbf{X}$ and Adjustment Variables $\mathbf{Z}$ from our N-D^{2}VD Estimator

$\mathbf{T} = \mathbf{T_{logit}}$							
		$p = 50$		$p = 100$		$p = 150$	
n		TPR	TNR	TPR	TNR	TPR	TNR
$n = 10000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	1.000	1.000
	$\mathbf{Z}$	1.000	0.994	1.000	0.94	1.000	0.999
$n = 20000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	1.000	1.000
	$\mathbf{Z}$	1.000	1.000	1.000	0.999	1.000	0.999
$\mathbf{T} = \mathbf{T_{missp}}$							
	n	$p = 50$		$p = 100$		$p = 150$	
n		TPR	TNR	TPR	TNR	TPR	TNR
$n = 10000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	1.000	1.000
	$\mathbf{Z}$	1.000	0.983	1.000	0.937	1.000	0.999
$n = 20000$	$\mathbf{X}$	1.000	1.000	1.000	1.000	1.000	1.000
	$\mathbf{Z}$	1.000	1.000	1.000	0.999	1.000	0.999

The closer to **1** *for TPR and TNR is better.*

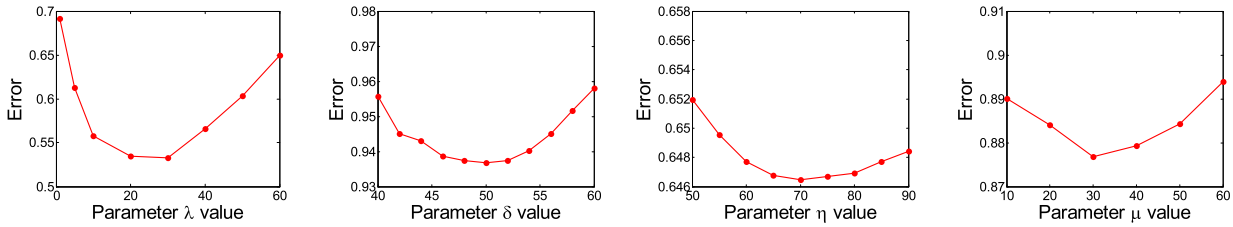


Fig. 3. Parameters tuning.

TABLE 6
The Top Ranked Features by Their Absolute ATE Estimated With our D²VD Estimator $\widehat{ATE}_{D^2VD}$, Comparing With the Baseline Estimator $\widehat{ATE}_{IPW}$ and $\widehat{ATE}_{DR}$

No.	Features	$\widehat{ATE}_{D^2VD}$ (SD)	$\widehat{ATE}_{IPW}$ (SD)	$\widehat{ATE}_{DR}$ (SD)	$ATE_{matching}$
1	No. friends ($>$ 166)	0.295 (0.018)	0.240 (0.026)	0.297(0.021)	0.276
2	Age ($>$ 33)	-0.284 (0.014)	-0.235 (0.029)	-0.302(0.068)	-0.263
3	Share Album to Strangers	0.229 (0.030)	0.236 (0.030)	-0.034(0.021)	n/a
4	With Online Payment	0.226 (0.019)	0.260 (0.029)	0.244(0.028)	n/a
5	With High-Definition Head Portrait	0.218 (0.028)	0.203 (0.032)	0.237(0.046)	n/a
6	With WeChat Album	0.191 (0.014)	0.237 (0.021)	0.097(0.050)	n/a
7	With Delicacy Plugin	0.124 (0.038)	-0.253 (0.037)	0.067(0.051)	0.099
8	Device (iOS)	0.100 (0.024)	0.206 (0.012)	0.060(0.021)	0.085
9	Add friends by Drift Bottle	-0.098 (0.012)	0.016 (0.019)	-0.115(0.015)	-0.032
10	Gender (Male)	-0.073 (0.017)	-0.240 (0.029)	0.065(0.055)	-0.097

The $ATE_{matching}$ is the "approximal ground truth" by matching method, "n/a" means that we cannot obtain the ATE from matching method since the number of matching samples are not sufficient.

estimate the propensity score more accurately for better treatment effect estimation. However, Table 4 show that our D²VD cannot separate the adjustment variables precisely, the value of TPR and TNR on adjustment variables $\mathbf{Z}$ is far from 1. This is induced by the incorrectly specified function between adjustment variables and outcome.

With considering the non-linear representation on adjustment variables $\mathbf{Z}$, our N-D²VD algorithm can simultaneously separate the confounders $\mathbf{X}$ and adjustment variables $\mathbf{Z}$ precisely, as shown in Table 5. Even in the setting $T = T_{missp}$, our algorithm can still simultaneously precisely separate the confounders and adjustment variables. This is the key reason that our N-D²VD algorithm achieves better performance than D²VD algorithm on treatment effect estimation. Hence, enabling us to estimate the ATE more accurately and with tighter confidence intervals than the state-of-the-art methods.

7.4 Experiments on Real Advertising Data

7.4.1 Dataset

The real online advertising dataset we used is collected during Sep. 2015 from Tencent WeChat App[5]. In WeChat, each user can share posts to his/her friends and receive posts from friends like in Twitter and Facebook. The advertisers can push advertisements to users, by merging them into a list of the user's wallposts. There are two types of feedback on the advertisements: "Like" and "Dislike".

The online advertising campaign is about LONGCHAMP handbags for young ladies[6]. This campaign contains 14,891 user feedbacks with Like and 93,108 Dislikes. For each user, we have 56 features including (1) demographic attributes, such as age, gender, (2) number of friends, (3) device (iOS or Android), and (4) the user settings on WeChat, for example, whether allowing strangers to see his/her album ("Share Album to Strangers") and whether installing the online payment service ("With Online Payment").

7.4.2 Experimental Settings

In our experiments, we set the feedback of users about the advertisement as the outcome Y. Specifically, we set the outcome $Y_i = 1$ when the user i likes the advertisement and $Y_i = 0$ if user i dislikes it. And we alternatively set one of the features as the treatment T and all other features as the variables $\mathbf{U}$. So that we can evaluate the ATE of each feature.

During the parameters tuning, we set the matching threshold $\epsilon = 5$, which makes the matching estimator is close to the exact matching and we can obtain the "approximal ground truth". With regard to parameters settings, we have λ and δ as the relative weights of L_1 norm on the confounder coefficients and adjustment variable coefficients, η as the relative weight of L_1 norm on average causal effect coefficients, and μ as the relative weight of L_2 norm for enforcing separation of confounders and adjustment variables. In detail, we tuned these 4 parameters with cross-validation by grid searching with the "approximal ground truth". In Fig. 3, we show the change of Bias $Error$ between estimated and true treatment effect with respect to different values of parameters. And we obtain the best parameters $\lambda = 30$, $\delta = 50$, $\eta = 70$ and $\mu = 30$ for our D²VD algorithm.

7.4.3 Treatment Effect Estimation

For each user feature, we employ our D²VD algorithm to estimate its ATE on the outcome. Table 6 shows the top-ranked features by their absolute ATE estimated with our

5. http://www.wechat.com/en/
6. http://en.longchamp.com/en/womens-bags

TABLE 7
Confounders and Adjustment Variables When we Set the "Add Friends by Shake" as Treatment

Confounders	Adjustment Variables
With Drift Bottle plugin	No. friends
Add friends by People Nearby	Age
Add friends by QQ Contacts	With WeChat Album
Without Friends Confirmation Plugin	Device

TABLE 8
Confounders and Adjustment Variables When we Set the "With WeChat Album" as Treatment

Confounders	Adjustment Variables
Open WeChat Album Service	Friends Count
With High-Definition Head Portrait	Age
Open to Strangers	With Drift Bottle Plugin
With Personal Information	Device

D^{2}VD estimator, comparing with baseline estimators and the "approximal ground truth" $ATE_{matching}$. Note that the $ATE_{matching}$ has very rigorous requirements on the sample size with exactly ma tching. For some user features, we do not have a sufficient number of samples thus we cannot derive their $ATE_{matching}$.

From Table 6, we have following observations.

O1. Our D^{2}VD estimator evaluates the ATE more accurately than baseline estimators. With separated confounders, the ATE estimated by our D^{2}VD estimator is closer to the "approximate ground truth" $ATE_{matching}$. While the IPW and DR estimators, which treat all variables as confounders, generate a huge error in estimating ATE for some features, even make the wrong estimation of the ATE polarity (positive or negative), such as feature $WithDelicacyPlugin$ for IPW estimator and feature $Gender$ for DR estimator.

O2. Our D^{2}VD estimator can reduce the variance of estimated ATE from baseline estimators. With regression on separated adjustment variables, our estimator obtains smaller SD than IPW and DR estimators, where the IPW estimator ignores the adjustment variables and the DR estimator makes regression on all variables, ignoring the variables separation.

O3. Younger ladies are with higher probability to like the advertisement about LONGCHAMP handbags. The ATE of $Age(>33)$ is -0.284 and $Gender(Male)$ is -0.073, which indicates that the younger ladies have a higher probability of like the advertisement. This is consistent with our intuition since the LONGCHAMP advertisement is mainly designed for young ladies as their potential customers.

Some features in Table 6 cannot be intuitively interpreted for their causal effect on users' Like behaviors. We attribute this to the mediational effect [32] of unobserved confounders or variables on them, which are common in previous studies in causal effects.

7.4.4 Case Studies on Variables Decomposition

We illustrate several case studies to demonstrate the separation results of confounders and adjustment variables when employing our D^{2}VD algorithm for causal inference in an online advertising campaign. Our case studies include two different treatments, "Add friends by Shake" and "With WeChat Album".

Treatment 1: *Add friends by Shake.* Shake[7] is a two-way function where both people using this function at the same time can see each other and make friends on WeChat. Table 7 shows the separation results between confounders and adjusted variables when we set "Add friends by Shake" as the treatment. The confounders are many other ways for adding friends on WeChat, which have a significant causal effect to the treatment, such as "With Drift Bottle Plugin" which is a channel to add friends via "Drift Bottle" among strangers, "Add friends by People Nearby" which allows one can look around and make friends with people nearby and "Add friends by QQ Contacts" which allows one make friends by contracts of QQ app (QQ is another communicational app of Tencent). The confounder "Without Friends Confirmation Plugin" which allows strangers to make friends with you without sending an initial friend request to you, indicating that you are willing to make friends with strangers. The confounders are precisely separated, hence we can better eliminate their confounding effect on the treatment and obtain more accurate ATE of the feature "Add friends by Shake".

While the adjustment variables, for example, the "No. friends" and "Age", are not associated with the treatment but have a significant effect on the outcome. As shown in Table 6, the features "No. friends" and "Age" are the top-2 ranked features that are causally related to outcome. With the precisely separated adjusted variables, we can reduce the variance of estimated ATE by our estimator.

Treatment 2: *With WeChat Album.* On WeChat, users can share albums with their friends. Table 8 shows the separation results of confounders and adjusted variables when we set "With WeChat Album" as treatment. One wants to have the album on WeChat, he/she need to open the service first, hence the feature "Open WeChat Album Service" should be the confounder. On the other hand, the users with a high-definition head portrait on WeChat may have an album with high probability. Therefore, the separated confounders are highly related to the treatment. While the adjusted variables are independent with treatment and have an important impact on the outcome, such as the "No. friends" and "Age". With the precisely separated confounders and adjustment variables of treatment "With WeChat Album", we can estimate its ATE more accurately with a smaller variance than other baseline estimators.

7.4.5 Like or Dislike Prediction

In addition to estimating the causal effect, we are also interested in whether the top k features selected by our adjusted estimator, can get good performance in predicting the Like and dislike behaviors of users. We compare with both IPW estimator, direct estimator, and the commonly used methods for correlation-based feature selection, including MRel (Maximum Relevance) [33] and mRMR (Maximum Relevance Minimum Redundancy) [34]. We use MAE as the

7. https://rumorscity.com/2014/07/25/how-to-add-friends-on-wechat-7-ways/

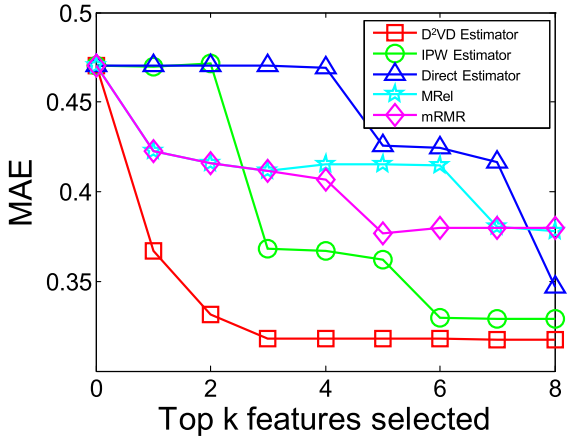


Fig. 4. Our D²VD estimator outperforms the baseline methods when selecting the top k significant causal features to predict whether user will like or dislike an advertisement.

evaluation metric, which is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|,$$

where n is the number of users in test data, $\hat{Y}_i$ and Y_i represent the predict and actual feedback of user i on the advertisement.

Fig. 4 shows that our adjusted estimator achieves the best prediction accuracy with a different number of features. Also, our method can get nearby the optimal prediction performance with much fewer features than other baselines. For example, when $k = 3$, the MAE value of our D²VD estimator (0.3178) is much smaller than those values of IPW estimator (0.3682), direct estimator (0.4701), MRel (0.4116) and mRMR (0.4115). Another important observation is that the two commonly used correlation-based feature selection methods perform worse than our method and even the IPW causal estimators. This demonstrates that the explainability brought by causal analysis can also significantly help to improve the prediction performances, as long as the confounding problems are subtly addressed.

7.5 Experiments on LaLonde Dataset

7.5.1 LaLonde Dataset

LaLonde [35] dataset[8] is a canonical benchmark in the causal inference literature [36], [37]. The LaLonde dataset used in our paper consists of two parts. The first part comes from a randomized experiment on a large scale job training program, the National Support Work Demonstration (NSW).[9] In the second part data, as [37] did, we replace the control group in a randomized experiment with another control group drawn from the Current Population Survey-Social Security Administration file (CPS-1) where the measured covariates are the same with the experimental data. The treatment variable in this data is whether the participant attends the particular job training program or not, and the outcome is earning in the year 1978. The data contains 10 raw observed variables, including earnings and employment status for the years 1974 and 1975, education status (years of schooling and an indicator for completed high school degree), age, ethnicity (indicators for black and Hispanic) and the married status.

Overall, there are *185* program participants (the treated units) and *260* nonparticipants (the control units) in the experimental data NSW. In the observational data CPS-1, we have *185* program participants and *15,992* nonparticipants. The randomized experimental data NSW provide the ground truth for estimating the treatment effect of the program. We estimate the treatment effect with the observational data CPS-1, comparing our proposed algorithm with the baselines.

TABLE 9
Results of ATE Estimation on LaLonde Dataset, Where the True Treatment Effect From Randomized Experiment is 1,794

Estimator	*Bias*	*SD*	*MAE*	*RMSE*
$\widehat{ATT}_{dir}$	10299.92	65.85	10299.92	10300.13
$\widehat{ATT}_{IPW}$	761.86	240.71	761.86	798.98)
$\widehat{ATT}_{DR}$	529.78	235.95	530.61	579.95
$\widehat{ATT}_{D^2VD}$	417.7	263.2	431.4	493.7
$\widehat{ATT}_{N-D^2VD}$	**370.6**	263.0	**400.2**	**454.5**

The smaller Bias and SD, the better.

7.5.2 Experimental Settings and Results

In our experiments, we randomly split the observational data CPS-1 as 6 partitions, with the first 3 partitions, we train our model and baseline models for parameters tuning with cross validation by grid searching, and test model performance and robustness with the last 3 partitions.

We report the results in Table 9, where the smaller $Bias$, SD, MAE, and $RMSE$, the better. From the results, we have the following observations. (1) Directly estimator failed due to the existence of confounding bias in the LaLonde data. (2) IPW estimator generates a big error on treatment effect estimation. The main reason is that the specification model of IPW is incorrect and the sample size between treated and control units is unbalanced. (3) By combining the IPW and regression model, the DR estimator achieves better performance than the IPW estimator. (3) Our proposed D²VD estimator achieves better performance compared with the baselines since D²VD can precisely separate the confounders and adjustment variables. (4) With considering the nonlinear presentation of adjustment variables, our proposed N-D²VD algorithm obtains better performance than our D²VD algorithm.

8 Conclusion

In this paper, we focus on how to evaluate the average treatment effect in a more precise way with tighter confidence intervals in observational studies. We argued that most previous causal methods based on propensity score are deficient because they usually treat all variables as confounders. Based on our causal diagram, we proposed to separate the confounders and adjustment variables from all observed variables for causal inference in the hope of reducing variance and

8. The dataset is available at http://users.nber.org/~rdehejia/data/nswdata2.html

9. Notice that we focus on the Dehejia and Wahha sampled dataset of the LaLonde.

improving the accuracy of average treatment effect estimation. Aiming at this, we proposed a Data-Driven Variable Decomposition (D^2VD) algorithm to jointly optimize the variables decomposition and ATE estimation. Moreover, we proposed a Non-linear D^2VD (N-D^2VD) algorithm to address the challenges of high-dimensional and nonlinear in observational studies. Theoretically, we proved that our algorithms can unbiased estimate the treatment effect and achieve lower variance than traditional methods. Experimental results on both synthetic data and real-world data verify the practical usefulness of our proposed models and the effectiveness of our proposed algorithms for ATE estimation in observational studies.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2018AAA0101900, No. 2018AAA0102004), National Natural Science Foundation of China (No. U1936219, No. 61772304, No. 61531006, No. U1611461), Beijing Academy of Artificial Intelligence (BAAI), Fundamental Research Funds for the Central Universities. Bo Li's research was supported by the Tsinghua University Initiative Scientific Research Grant, No. 2019THZWJC11; National Natural Science Foundation of China, No. 71490723 and No. 71432004; Science Foundation of Ministry of Education of China, No. 16JJD630006. Fei Wu's research was supported by National Science Foundation for Distinguished Young Scholars No. 61625107 and National Natural Science Foundation of China No. 61751209.

References

[1] K. Kuang *et al.*, "Causal inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.
[2] P. W. Holland, "Statistics and causal inference," *J. Amer. Stat. Assoc.*, vol. 81, pp. 945–960, 1986.
[3] R. Lewis and D. Reiley, "Retail advertising works! measuring the effects of advertising on sales via a controlled experiment on yahoo!" Yahoo Research Technical Report, 2009.
[4] R. Kohavi and R. Longbotham, "Unexpected results in online controlled experiments," *ACM SIGKDD Explorations Newslett.*, ACM New York, NY, USA, vol. 12, no. 2, pp. 31–35, 2011.
[5] L. Bottou *et al.*, "Counterfactual reasoning and learning systems: The example of computational advertising," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3207–3260, 2013.
[6] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[7] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Res.*, vol. 46, no. 3, pp. 399–424, 2011.
[8] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study," *Stat. Med.*, vol. 23, no. 19, pp. 2937–2960, 2004.
[9] K. Kuang, M. Jiang, P. Cui, and S. Yang, "Steering social media promotion with effect strategies," in *Proc. IEEE 16th Int. Conf. Data Minin*, 2016, pp. 985–990.
[10] M. Á. Hernán, B. Brumback, and J. M. Robins, "Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men," *Epidemiology*, vol. 11, no. 5, pp. 561–570, 2000.
[11] M. A. Hernán, B. A. Brumback, and J. M. Robins, "Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures," *Stat. Med.*, vol. 21, pp. 1689–1709, 2002.
[12] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.
[13] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statist. Sci.: A Review J. Inst. Math. Statist.*, vol. 25, no. 1, 2010, Art. no. 1.
[14] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, "Doubly robust estimation of causal effects," *Amer. J. Epidemiol.*, vol. 173, no. 7, pp. 761–767, 2011.
[15] V. L. Dos Reis and A. Culotta, "Using matched samples to estimate the effects of exercise on mental health from twitter," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 182–188.
[16] M. Lechner, "Earnings and employment effects of continuous gff-the-job training in east germany after unification," *J. Bus. Econ. Statist.*, vol. 17, no. 1, pp. 74–90, 1999.
[17] W. Sun, P. Wang, D. Yin, J. Yang, and Y. Chang, "Causal inference via sparse additive models with application to online advertising," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 297–303.
[18] M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer, "Variable selection for propensity score models," *Amer. J. Epidemiol.*, vol. 163, no. 12, pp. 1149–1156, 2006.
[19] T. J. VanderWeele and I. Shpitser, "A new criterion for confounder selection," *Biometrics*, vol. 67, no. 4, pp. 1406–1413, 2011.
[20] B. C. Sauer, M. A. Brookhart, J. Roy, and T. VanderWeele, "A review of covariate selection for non-experimental comparative effectiveness research," *Pharmacoepidemiol. Drug Safety*, vol. 22, pp. 1139–1145, 2013.
[21] A. Bloniarz, H. Liu, C.-H. Zhang, J. Sekhon, and B. Yu, "Lasso adjustments of treatment effect estimates in randomized experiments," *Proc. Nat. Acad. Sci. USA*, vol. 113, pp. 7383–7390, 2016.
[22] B. K. Lee, J. Lessler, and E. A. Stuart, "Improving propensity score weighting using machine learning," *Statist. Med.*, vol. 29, no. 3, pp. 337–346, 2010.
[23] X. Su, J. Kang, J. Fan, R. A. Levine, and X. Yan, "Facilitating score and causal inference trees for large observational studies," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2955–2994, 2012.
[24] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert, "Evaluating online ad campaigns in a pipeline: Causal models at scale," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 7–16.
[25] P. R. Rosenbaum, "Model-based direct adjustment," *J. Amer. Stat. Assoc.*, vol. 82, pp. 387–394, 1987.
[26] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1097–1104.
[27] A. Basu, D. Polsky, and W. G. Manning, "Use of propensity scores in non-linear response models: the case for health care expenditures," National Bureau Economic Res., Cambridge, MA, 2008.
[28] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, and F. Wang, "Treatment effect estimation with data-driven variable decomposition," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 140–146.
[29] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 27, pp. 7353–7360, 2016.
[30] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.
[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[32] D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets, "A comparison of methods to test mediation and other intervening variable effects," *Psychol. Methods*, vol. 7, no. 1, 2002, Art. no. 83.
[33] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, 2003.
[34] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[35] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *Amer. Econ. Rev.*, vol. 76, pp. 604–620, 1986.
[36] A. Diamond and J. S. Sekhon, "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Rev. Econ. Stat.*, vol. 95, no. 3, pp. 932–945, 2013.
[37] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Anal.*, vol. 20, no. 1, pp. 25–46, 2012.

Kun Kuang received the PhD degree from Tsinghua University, in 2019. He is currently an assistant professor with the College of Computer Science and Technology, Zhejiang University. His main research interests including causal inference and causally regularized machine learning. He was a visiting scholar at Stanford University. He has published more than 10 papers in major international journals and conferences, including SIGKDD, ICML, ACM MM, AAAI, the *ACM Transactions on Knowledge Discovery from Data*, and ICDM, etc.

Peng Cui received the PhD degree in computer science from Tsinghua University, in 2010, and he is an assistant professor with Tsinghua. He has vast research interests in data mining, multimedia processing, and social network analysis. Until now, he has published more than 60 papers in conferences such as SIGIR, AAAI, ICDM, etc. and journals such as the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Image Processing*, *Data Mining and Knowledge Discovery*, etc. He won five best paper awards in recent four years, including ICDM2015 Best Student Paper Award, ICME 2014 Best Paper Award, etc. In 2015, he was awarded as ACM China Rising Star. Now his research is sponsored by National Science Foundation of China, Samsung, Tencent, etc. He also serves as guest editor, co-ehair, PC member, and Reviewer of several high-level international conferences, workshops, and journals.

Hao Zou received the BE degree from the Department of Computer Science and Technology, Tsinghua University, in 2018. He is currently working toward the PhD degree in the Department of Computer Science and Technology, Tsinghua University. His main research interests including causal inference, counterfactual learning, and high dimensional inference.

Bo Li received the bachelor's degree in mathematics from Peking University, and the PhD degree in statistics from the University of California, Berkeley. He is an associate professor with the School of Economics and Management, Tsinghua University. His research interests include statistical methods for high-dimensional data, statistical causal inference and data-driven decision making. He has published widely in academic journals across a range of fields including statistics, management science and economics.

Jianrong Tao received the BS degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and the MS degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2017. He is currently working ar NetEase Fuxi AI Lab, Hangzhou, China. His research interests include data mining, machine learning, network analysis, and user profiling.

Fei Wu received the PhD degree from the College of Computer Science, Zhejiang University. Now he is the professor and dean of the College of Computer Science, Zhejiang University. His main research interests include multimedia information analysis and retrieval, digital library.

Shiqiang Yang received the BE and ME degrees from the Department of Computer Science and Technology, Tsinghua University, in 1977 and 1983, respectively. He is now a professor with Tsinghua University. His research interests include multimedia technology and systems, video compression and streaming, content-based retrieval for multimedia information, multimedia content security, and digital right management. He has published more than 100 papers and MPEG standard proposals. Professor Yang has organized many conferences as program chair or TPC member including PCM05, PCM06 Workshop On ACM Multimedia05, MMM06, ICME06, MMSP05, ASWC06, etc.