Treatment Effect Estimation with Data-Driven Variable Decomposition

Kun Kuang[1,2], **Peng Cui**[1,2], **Bo Li**[3], **Meng Jiang**[4], **Shiqiang Yang**[1,2], **Fei Wang**[5]
[1]Tsinghua National Laboratory for Information Science and Technology
[2]Department of Computer Science and Technology, Tsinghua University
[3]School of Economics and Management, Tsinghua University
[4]Department of Computer Science, University of Illinois Urbana-Champaign
[5]Department of Healthcare Policy and Research, Weill Cornell Medical School, Cornell University
kk14@mails.tsinghua.edu.cn, cuip@tsinghua.edu.cn, libo@sem.tsinghua.edu.cn
mjiang89@illinois.edu, yangshq@tsinghua.edu.cn, feiwang03@gmail.com

Abstract

One fundamental problem in causal inference is the treatment effect estimation in observational studies when variables are confounded. Control for confounding effect 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 big data era. In this paper, we 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 show experimentally that the proposed D^2VD algorithm can automatically separate the variables precisely, and estimate treatment effect more accurately and with tighter confidence intervals than the state-of-the-art methods on both synthetic data and real online advertising dataset.

Introduction

Causal inference, which refers to the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect (Holland 1986), is a powerful statistical modeling tool for explanatory analysis. The gold standard approaches for causal inference are randomized experiments, for example, A/B testing (Lewis and Reiley 2009), where different treatments are randomly assigned to units [1]. However, the fully randomized experiments are usually extremely expensive (Kohavi and Longbotham 2011) or sometimes even infeasible (Bottou et al. 2013) 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 1983) proposed a statistical framework for treatment effect estimation based on propensity score adjustment. Such framework has been widely used in observational causal study, including matching, stratification, inverse weighting and regression on propensity score (Austin 2011; Chan et al. 2010; Lunceford and Davidian 2004). The inverse propensity weighting is a most commonly used method and has been part of a large family of causal models known as marginal structural model (Hernán, Brumback, and Robins 2000; 2002). With combination of inverse propensity weighting and regression, (Bang and Robins 2005) proposed a doubly robust estimator. These methods have been widely used in various fields, including economics (Stuart 2010), epidemiology (Funk et al. 2011), health care (Reis et al. 2015), social science (Lechner 1999), political science (Dudík, Langford, and Li 2011) and advertising (Sun et al. 2015).

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. However, most of these works treat all observed variables as confounders when estimating 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 **Z** as shown in our causal diagram in Fig. 1. 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. (Brookhart et al. 2006; VanderWeele and Shpitser 2011) have advocated that the ad-

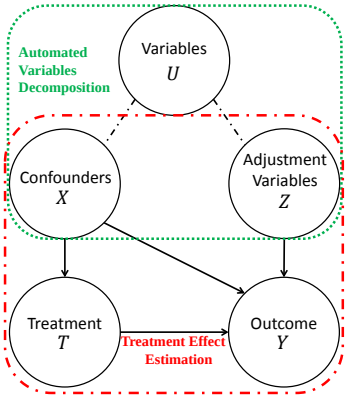

Figure 1: Our causal diagram. We separate all observed variables **U** into three different sets: (1) Confounders **X**, which are associated with the treatment T and may be causally related to the outcome Y, (2) Adjustment Variables **Z**, which are causally related to outcome Y, but independent with treatment T, and (3) Irrelevant Variables **I** (Omitted), which are independent with both treatment and outcome.

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

justment variables should be included in the causal inference model. And (Sauer et al. 2013) suggested that conditioning on such adjustment variables is unnecessary to remove bias but can reduce variance in treatment effect estimation. In randomized experiments setting, (Bloniarz et al. 2016) have proved that adjusting for the adjustment variables by lasso can reduce the variance of 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.

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 algorithm, where a regularized integrated regression model is presented to enable confounder separation and ATE estimation simultaneously.
- The advantages of D^2VD algorithm are demonstrated in both synthetic and real world data. It can also be straightforwardly applied into other causal inference studies, such as social marketing, health care and public policy.

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.

Notations and Assumptions

As described in our causal diagram in Fig.1, 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, that is $T = 1$, as treated units and the others with $T = 0$ as control units. Then for each unit indexed by $i = 1, 2, \cdots, m$, we observe a treatment T_i, an outcome Y_i^{obs} and a vector of variables U_i with dimension n. Our observed outcome Y_i^{obs} of unit i can be denoted by:

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

For any column vector $\mathbf{v} = (v_1, v_2, \cdots, v_m)^T$, let $\|\mathbf{v}\|_2^2 = \sum_{i=1}^m v_i^2$, and $\|\mathbf{v}\|_1 = \sum_{i=1}^m |v_i|$.

In observational studies, there are three standard assumptions (Rosenbaum and Rubin 1983) for ATE estimation.

Assumption 1: Stable Unit Treatment Value. The distribution of potential outcome 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 is independent of potential outcome when given the observed variables. Formally, $T \perp \big(Y(0), Y(1)\big) | \mathbf{U}$.

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

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:

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

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

The Eq. (2) is infeasible, because for each unit, we can only observe one potential outcome corresponding to its treatment status, treated or control. This is called "the counterfactual problem" (Chan et al. 2010).

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. However, in observational studies, comparing two samples directly is likely to have bias if the treatment assignment is not random, as confounding impact is not taken into account (Chan et al. 2010).

To unbiasedly evaluate the ATE in observational studies, one have to control the impact of confounders. Under the assumptions (1,2,3), (Rosenbaum and Rubin 1983) 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 with treated status ($T = 1$) of a unit given all variables $\mathbf{U}$. Actually, 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, (Rosenbaum 1987) 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 (Athey and Imbens 2016). The transformed outcome $Y^\star$ is defined as

$$Y^\star = Y^{obs} \cdot \tfrac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \tfrac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}, \tag{4}$$

and the IPW estimator is defined as

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

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 direct impact on outcome.

Therefore, based on our causal diagram as shown in Fig. 1, 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.1). And then, we propose a new adjusted estimator by incorporating adjustment variables to reduce the variance of estimated ATE under 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 $\mathbf{X}$ are confounders, $\mathbf{Z}$ are adjustment variables and $\mathbf{I}$ are irrelevant variables.

With assumption 4, we introduce our adjusted transformed outcome Y^+ based on the transformed outcome in Eq. (4) with the definition as

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

where $\phi(\mathbf{Z})$ helps to reduce the variance among Y, which are associated with $\mathbf{Z}$.

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

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

And our adjusted estimator has following properties.

• Firstly, under assumptions 1-4, our adjusted estimator $\widehat{ATE}_{adj}$ is unbiased, that is

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

This property is obvious with the Pearl's back-door criterion (Pearl 2009). Since the conditioning set $\mathbf{X}$ blocks all back door paths linking treatment T and outcome Y, while not contains any descendants of T in our causal diagram.

• Secondly, the asymptotic variance of our adjusted estimator $\widehat{ATE}_{adj}$ is no greater than IPW estimator $\widehat{ATE}_{IPW}$.

Comparing the Eq. (5) and (7), we know that the IPW estimator only considered the confounders $\mathbf{X}$ when provided all variables $\mathbf{U}$, while our estimator utilized the adjustment variables $\mathbf{Z}$ to make adjustments on outcome Y for reducing variance. The similar adjustments have been proved can reduce the variance of ATE estimation in randomized experiments by (Bloniarz et al. 2016).

Automated Variables Decomposition

D^{2}VD Algorithm

With Eq. (8), we can get $E(Y^+) = E(Y(1) - Y(0))$, and obtain the estimated ATE by regressing our adjusted transformed outcome Y^+ against the variables $\mathbf{U}$ and minimizing the following objective function

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2. \tag{9}$$

Then we can estimate the ATE by $\widehat{E}(h(\mathbf{U}))$ and get the conditional ATE by $h(\mathbf{U})$, or we can obtain the estimated ATE by setting the $h(\mathbf{U})$ as a constant in Eq. (9).

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

$$\phi(\mathbf{Z}) = \mathbf{Z}\alpha, \quad h(\mathbf{U}) = \mathbf{U}\gamma, \tag{10}$$

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

$$e(\mathbf{X}) = p(T = 1|\mathbf{X}) = \tfrac{1}{1 + \exp(-\mathbf{X}\beta)}. \tag{11}$$

In the specifications of Eq. (10, 11), we have assumed the knowledge of the variables 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 observed variables $\mathbf{U}$ to replace $\mathbf{X}$ and $\mathbf{Z}$ instead, and propose a data-driven approach to automatically separate confounders and adjustment variables. We update our objective function in Eq. (9) as:

$$\begin{aligned} minimize \quad & \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \\ s.t. \quad & \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau, \\ & \|\alpha\|_1 \le \lambda,\ \|\beta\|_1 \le \delta,\ \|\gamma\|_1 \le \eta,\ \|\alpha \odot \beta\|_2^2 = 0. \end{aligned} \tag{12}$$

where $W(\beta) := \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))}$ and $\sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta))$ represents the loss function when estimating the propensity score. And $\odot$ refers to Hadamard product. With the formula $\|\alpha \odot \beta\|_2^2 = 0$, the coefficient vector α is optimized for separating the adjustment variables $\mathbf{Z}$ and β is for separating confounders $\mathbf{X}$ form variables $\mathbf{U}$.

In particular, we employ an orthogonal regularizer on α and β to ensure the separation of confounders and adjustment variables. In addition, we add $L1$ penalties on α, β and γ to eliminate irrelevant variables $\mathbf{I}$ to further reduce variance and address the sparseness problem of variables.

These lead to the following optimization problem, which is to minimize $\mathcal{J}(\alpha, \beta, \gamma)$.

$$\begin{aligned} \mathcal{J}(\alpha, \beta, \gamma) &= f(\alpha, \beta, \gamma) + g(\alpha, \beta, \gamma), \\ f(\alpha, \beta, \gamma) &= \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2 + \mu\|\alpha \odot \beta\|_2^2 \\ &\quad + \tau \textstyle\sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)), \\ g(\alpha, \beta, \gamma) &= \lambda\|\alpha\|_1 + \delta\|\beta\|_1 + \eta\|\gamma\|_1. \end{aligned} \tag{13}$$

With the operator splitting property of proximal gradient algorithm (Parikh and Boyd 2013), we can get the optimized parameter (i.e., $\alpha^{(t+1)}$) at the t^{th} iteration by proximal operator $prox_{\kappa^{(t)}g}$ of function $g(\cdot)$ with the step size $\kappa^{(t)}$:

$$\alpha^{(t+1)} = prox_{\kappa^{(t)}g}\left(\alpha^{(t)} - \kappa^{(t)} \tfrac{\partial f(\cdot)}{\partial \alpha}\right) \tag{14}$$

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

$$prox_{\kappa^{(t)}g}(x) = \begin{cases} x_i - \kappa^{(t)} \cdot \lambda & x_i \ge \kappa^{(t)} \cdot \lambda \\ 0 & |x_i| \le \kappa^{(t)} \cdot \lambda \\ x_i + \kappa \cdot \lambda & x_i \le -\kappa^{(t)} \cdot \lambda \end{cases} \tag{15}$$

The λ in Eq. (15) is the coefficient of parameter α in function $g(\cdot)$. If the optimized parameter is β, then it should be δ and it should be η for optimizing parameter γ.

With the proximal gradient algorithm, we can minimize the objective function in Eq. (13). That is, starting from some random initialization on α, β, γ, we solve each them alternatively with the other two parameters as fixed and step by step until convergence. Our Data-Driven Variable Decomposition algorithm is described in Algorithm 1.

During each iteration in Algorithm 1, we update the parameters α, β and γ with OPTIMIZATION as described in Algorithm 2, where the function $\hat{f}_\kappa(\cdot)$ is defined as:

$$\hat{f}_\kappa(x, y) = f(y) + (x - y)\tfrac{\partial f(\cdot)}{\partial x}^{\mathrm{T}} + (1/(2\kappa))\|x - y\|_2^2. \tag{16}$$

And the gradients of the function $f(\alpha, \beta, \gamma)$ with the respect

Algorithm 1 Data-Driven Variable Decomposition (D^2VD)

Require: Initialization $\mathcal{J}^{(0)} = \mathcal{J}(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})$.
Ensure: $\mathcal{J}^{(0)} \geq 0$, $\mathcal{J}^{(t+1)} < \mathcal{J}^{(t)}$
1: **for** $t = 0, 1, 2, \cdots$ **do**
2: Calculate $\frac{\partial f(\cdot)}{\partial \alpha}$, $\frac{\partial f(\cdot)}{\partial \beta}$ and $\frac{\partial f(\cdot)}{\partial \gamma}$
3: $\alpha^{(t+1)} = \text{OPTIMIZATION}(\alpha, t)$
4: $\beta^{(t+1)} = \text{OPTIMIZATION}(\beta, t)$
5: $\gamma^{(t+1)} = \text{OPTIMIZATION}(\gamma, t)$
6: $\mathcal{J}^{(t+1)} = \mathcal{J}(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})$
7: **end for**

to the variables (α, β, γ) are:

$$\begin{aligned}
\frac{\partial f(\cdot)}{\partial \alpha} &= -2(W(\beta)\cdot \mathbf{1}^{\mathrm{T}} \odot \mathbf{U})^{\mathrm{T}} \cdot \mathbf{R} + 2\mu\alpha \odot \beta \odot \beta, \\
\frac{\partial f(\cdot)}{\partial \beta} &= 2\left((Y - \mathbf{U}\alpha)\cdot \mathbf{1}^{\mathrm{T}} \odot \tfrac{\partial W(\beta)}{\partial \beta}\right)^{\mathrm{T}} \cdot \mathbf{R} \\
&\quad + \mathbf{U}^T(T - \exp(\mathbf{U}\beta)) + 2\mu\alpha \odot \beta \odot \alpha, \\
\frac{\partial f(\cdot)}{\partial \gamma} &= -2\mathbf{U}^{\mathrm{T}} \cdot \mathbf{R}.
\end{aligned}$$

where $\mathbf{R} = \left((Y - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\right)$ and $\frac{\partial W(\beta)}{\partial \beta} = (2T - \mathbf{1}) \odot \exp\left((\mathbf{1} - 2T) \odot \mathbf{U}\beta\right) \odot (\mathbf{1} - 2T) \cdot \mathbf{1}^{\mathrm{T}} \odot \mathbf{U}$.

With the optimized parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ by Algorithm 1, we can separate the confounders as $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$, adjustment variables as $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$ and estimate the ATE as $\widehat{ATE}_{D^2VD} = E(\mathbf{U}\hat{\gamma})$.

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

Complexity Analysis

The complexity of our D^2VD algorithm is dominated by the step of calculating the gradients of function $f(\alpha, \beta, \gamma)$ with respect to the variables. The complexity of $\frac{\partial f(\cdot)}{\partial \alpha}$, $\frac{\partial f(\cdot)}{\partial \beta}$ and $\frac{\partial f(\cdot)}{\partial \gamma}$ are all $O(mn)$, where m is the sample size and n is the dimension of all observed variables. With considering that only constant time operations is involved in the for-loop and while-loop in our algorithms, therefore, the complexity of our D^2VD algorithm is $O(mn)$.

Parameters Tuning

The main challenge of parameters tuning for ATE estimation in observational studies is that there is no ground truth about the true ATE in practice.

To address this challenge, we employed the matching method to evaluate the ATE and set it as "approximal ground truth" like (Athey and Imbens 2016) did. Specifically, for each unit i, find its closest match among the units with opposite treatment status:

$$match(i) = \arg\min_{j:T_j = 1 - T_i} \|U_i - U_j\|_2^2. \tag{17}$$

We drop unit i if $match(i) > \epsilon$, that makes the matching approximate to exactly matching. We can estimate ATE with the matching estimator by comparing the average outcome between the matched treated and control units sets, and set it as "approximal ground truth", denoted by $ATE_{matching}$.

Algorithm 2 OPTIMIZATION(o, t)

1: Set $\kappa = 1$
2: **while** 1 **do**
3: Let $o^{(t+1)} = prox_{\kappa g}\left(o^{(t)} - \kappa \frac{\partial f(\cdot)}{\partial o}\right)$
4: **break if** $f(o^{(t+1)}) \leq \hat{f}_\kappa(o^{(t+1)}, o^{(t)})$
5: Update $\kappa = \frac{1}{2}\kappa$
6: **end while**
7: **return** o^{t+1}

With the "approximal ground truth", we can tune parameters of our algorithm with cross validation.

Experiments

We apply our algorithm on the synthetic dataset and real online advertising dataset to estimate the ATE.

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.
- *IPW Estimator* $\widehat{ATE}_{IPW}$ (Rosenbaum and Rubin 1983): 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* $\widehat{ATE}_{DR}$ (Bang and Robins 2005): It evaluates the ATE by combination of IPW and regression methods. It ignores the separation of confounders and adjustment variables.
- *Non-Separation Estimator* $\widehat{ATE}_{D^2VD(-)}$: It's a weakened version of our D^2VD estimator. It has no variables separation step by setting coefficient $\mu = 0$ in Eq. (13).

In this paper, we implemented $\widehat{ATE}_{IPW}$ and $\widehat{ATE}_{DR}$ with **lasso regression** for variables selection. The difference between $\widehat{ATE}_{DR}$ and $\widehat{ATE}_{D^2VD(-)}$ is that the former estimates ATE sequentially but the latter does with joint optimization.

Experiments on Synthetic Data

Dataset To generate the synthetic dataset, we set the sample size $m = \{1000, 5000\}$ and the dimension of observed variables $n = \{50, 100, 200\}$. We first generate the variables $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I}) = (\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i})$ with independent gaussian distributions as

$$\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i} \overset{iid}{\sim} \mathcal{N}(0, 1),$$

where n_x, n_z and n_i represent the dimension of confounders $\mathbf{X}$, adjustment variables $\mathbf{Z}$ and irrelevant variables $\mathbf{I}$, respectively. And $n_x = 0.2 * n$, $n_z = 0.2 * n$, $n_i = 0.6 * n$.

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

$T_{logit} \sim Bernoulli(1/(1 + \exp(-\sum_{i=1}^{n_x} x_i)))$ and
$T_{missp} = 1\ if\ \sum_{i=1}^{n_x} x_i > 0.5,\ T_{missp} = 0\ otherwise.$

The outcome Y is generated as

$$Y = \sum_{j=\frac{n_x}{2}}^{n_x} \mathbf{x}_j \cdot \omega_j + \sum_{j=1}^{n_z} \mathbf{z}_k \cdot \rho_k + T + \mathcal{N}(0, 2),$$

Table 1: Results on synthetic dataset: the true ATE is **1**. The $Bias$ refers to the absolute error between the true and estimated ATE, that is $Bias = |\widehat{ATE} - ATE|$. SD, MAE and RMSE represent the standard deviations, mean absolute errors and root mean square errors of $\widehat{ATE}$ after 50 times independently experiments, respectively.

	n	$n = 50$				$n = 100$				$n = 200$			
T/m	Estimator	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE
$T = T_{logit}$ $m = 1000$	$\widehat{ATE}_{dir}$	0.418	0.409	0.479	0.582	0.302	0.490	0.472	0.571	0.405	0.628	0.574	0.720
	$\widehat{ATE}_{IPW} + lasso$	0.078	0.310	0.252	0.317	0.097	0.356	0.295	0.366	0.073	0.328	0.267	0.320
	$\widehat{ATE}_{DR} + lasso$	0.060	0.181	0.152	0.189	0.067	0.190	0.155	0.199	0.081	0.181	0.169	0.190
	$\widehat{ATE}_{D^2VD(-)}$	0.053	0.138	0.124	0.146	0.064	0.130	0.117	0.144	**0.018**	0.170	0.128	0.162
	$\widehat{ATE}_{D^2VD}$	**0.045**	**0.108**	**0.091**	**0.116**	**0.019**	**0.114**	**0.093**	**0.115**	0.067	**0.144**	**0.130**	**0.152**
$T = T_{logit}$ $m = 5000$	$\widehat{ATE}_{dir}$	0.418	0.170	0.418	0.451	0.659	0.181	0.659	0.681	0.523	0.412	0.555	0.653
	$\widehat{ATE}_{IPW} + lasso$	0.036	0.201	0.163	0.202	0.034	0.222	0.194	0.213	**0.032**	0.341	0.274	0.325
	$\widehat{ATE}_{DR} + lasso$	0.051	0.079	0.071	0.094	0.106	0.075	0.114	0.127	0.055	0.084	0.086	0.096
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.080	0.118	0.137	0.114	0.102	0.121	0.150	0.164	0.076	0.164	0.179
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.072**	**0.061**	**0.078**	**0.023**	**0.073**	**0.061**	**0.073**	0.042	**0.068**	**0.062**	**0.076**
$T = T_{missp}$ $m = 1000$	$\widehat{ATE}_{dir}$	0.664	0.387	0.670	0.766	0.273	0.445	0.436	0.518	0.380	0.766	0.691	0.848
	$\widehat{ATE}_{IPW} + lasso$	0.266	0.279	0.319	0.384	0.298	0.295	0.328	0.417	0.191	0.482	0.403	0.514
	$\widehat{ATE}_{DR} + lasso$	0.138	0.187	0.174	0.231	0.253	0.197	0.269	0.320	**0.050**	0.218	0.170	0.222
	$\widehat{ATE}_{D^2VD(-)}$	0.269	0.162	0.270	0.313	0.129	0.162	0.170	0.206	0.175	0.207	0.236	0.269
	$\widehat{ATE}_{D^2VD}$	**0.066**	**0.113**	**0.102**	**0.129**	**0.019**	**0.119**	**0.101**	**0.120**	0.059	**0.177**	**0.149**	**0.184**
$T = T_{missp}$ $m = 5000$	$\widehat{ATE}_{dir}$	0.446	0.180	0.446	0.480	0.587	0.323	0.587	0.662	0.778	0.246	0.778	0.812
	$\widehat{ATE}_{IPW} + lasso$	0.148	0.133	0.161	0.198	0.172	0.167	0.199	0.239	0.142	0.224	0.206	0.263
	$\widehat{ATE}_{DR} + lasso$	0.119	0.073	0.123	0.139	0.100	0.067	0.107	0.120	0.127	0.079	0.127	0.148
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.070	0.119	0.132	0.058	**0.067**	0.069	0.086	0.068	0.055	0.073	0.086
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.055**	**0.052**	**0.063**	**0.039**	0.068	**0.066**	**0.075**	**0.032**	**0.047**	**0.049**	**0.055**

In this dataset, the features $(\mathbf{x}_{\frac{n_x}{2}}, \mathbf{x}_{\frac{n_x}{2}+1}, \cdots, \mathbf{x}_{n_x})$ are correlated to the treatment and outcome, simulating a confounding effect. The *true treatment effect* in this dataset is **1**.

ATE 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 Tab.1, where the smaller $Bias$, SD, MAE and RMSE are better. From Tab.1, we have following observations.

First, the direct estimator is failed (with large $Bias$) under different settings because it did not consider the confounding effect. Second, the IPW estimator can unbiasedly (with small $Bias$) estimate the ATE when $T = T_{logit}$, but with a big $Bias$ when propensity score model is misspecified by setting $T = T_{missp}$. With combination of IPW and regression models, DR estimator can get better performance than IPW estimator, especially when $T = T_{missp}$. Third, our $D^2VD(-)$ estimator, which has no variables separation step, can get the similar results with DR estimator. But with considering the separation between confounders and adjustment variables, our D^2VD estimator can improve the accuracy (smaller $Bias$) and reduce the variance (smaller SD) for ATE estimation from $D^2VD(-)$, DR and other baseline estimators under different settings.

Variables Decomposition As we described before, with the optimized $\hat{\alpha}$ and $\hat{\beta}$, our 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 record the true positive rate (TPR) and true negative rate (TNR) in Tab. 2. The formulations of TPR and TNR for separated confounders **X** are defined as

$$\text{TPR} = \frac{\#\{\hat{\beta}_i \neq 0, \beta_i \neq 0\}}{\#\{\beta_i \neq 0\}}, \text{TNR} = \frac{\#\{\hat{\beta}_i = 0, \beta_i = 0\}}{\#\{\beta_i = 0\}}. \quad (18)$$

Table 2: Separation results of confounders **X** and adjustment variables **Z**. The closer to **1** for TPR and TNR is better.

$\mathbf{T} = \mathbf{T_{logit}}$							
		$n = 50$		$n = 100$		$n = 200$	
m		TPR	TNR	TPR	TNR	TPR	TNR
$m = 1000$	**X**	1.000	0.917	0.977	0.948	0.966	0.906
	Z	1.000	0.973	1.000	0.983	1.000	0.984
$m = 5000$	**X**	1.000	0.923	1.000	0.887	0.994	0.989
	Z	1.000	0.975	1.000	0.987	1.000	0.994
$\mathbf{T} = \mathbf{T_{missp}}$							
$m = 1000$	**X**	1.000	0.844	0.997	0.866	0.867	0.977
	Z	1.000	0.982	1.000	0.987	1.000	0.983
$m = 5000$	**X**	1.000	0.843	1.000	0.837	0.998	0.965
	Z	1.000	0.986	1.000	0.990	1.000	0.994

In the same way, we calculate the TPR and TNR for separated adjustment variables **Z** via Eq. (18) by using α.

Tab. 2 shows that our D^{2}VD algorithm can separate the confounders **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 if setting $T = T_{missp}$, our algorithm can still precisely separate the confounders and adjustment variables. This enables us to estimate the ATE more accurately and with tighter confidence intervals than the state-of-the-art methods.

Experiments on Real World data

Dataset The real online advertising dataset we used is collected during Sep. 2015 from Tencent WeChat App[2]. In WeChat, each user can share posts to his/her friends and receive posts from friends as like in the Twitter and Facebook. The advertisers can push advertisements to users, by merging them into 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

[2]http://www.wechat.com/en/

Table 3: The top ranked features by their absolute ATE estimated with our D^{2}VD estimator $\widehat{ATE}_{D^2VD}$, comparing with the baseline estimator $\widehat{ATE}_{IPW}$ and $\widehat{ATE}_{DR}$. 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.

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

handbags for young ladies[3]. 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").

Experimental Settings In our experiments, we set the feedback of users about the advertisement as 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 **U**. So that we can evaluate the ATE of each feature.

During the parameters tuning, we set the matching threshold $\epsilon = 5$, which make the matching estimator is close to the exactly matching. The hyper-parameters of λ, δ, τ, η and μ set as 30, 50, 90, 70 and 30 by using grid search.

ATE Estimation For each user feature, we employ our D^{2}VD algorithm to estimate its ATE on the outcome. Tab. 3 shows the top ranked features by their absolute ATE estimated with our 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 matching. For some user features, we do not have a sufficient number of samples thus we cannot derive their $ATE_{matching}$.

From Tab. 3, we have following observations.

O1. Our D^{2}VD estimator evaluate 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 huge error in estimating ATE for some features, even make wrong estimation of the ATE polarity (positive of 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 obtain smaller SD than IPW and DR estimators, where IPW estimator ignores the adjustment variables and 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 indicate that the younger ladies have higher probability to like the advertisement. This is consistent with our intuition since the LONGCHAMP advertisement is mainly designed for young ladies as their potential customers.

[3]http://en.longchamp.com/en/womens-bags

Table 4: Confounders and adjusted variables when we set feature "Add friends by Shake" as treatment.

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

Variables Decomposition Tab. 4 shows the separation results between confounders and adjusted variables when we set feature "Add friends by Shake" as the treatment. Shake[4] is a two way function where both people using this function at the same time can see each other and make friends on WeChat. In Tab. 4, the confounders are many other ways for adding friends on WeChat, indicating the separated confounders have significant causal association with treatment. While the adjustment variables, for example, the "No. friends" and "Age", are not associated with treatment but have significant effect on outcome, as shown in Tab. 3, they are the top ranked features.

The results demonstrate that our proposed D^2VD algorithm can precisely separate the confounders and adjustment variables in practical. With the separated confounders, our estimator can obtain an accurate ATE, and reduce the variance of estimated ATE by the adjustment variables.

Conclusion

In this paper, we focus on how to evaluate the average treatment effect in a more precisely way with tighter confidence intervals in observational studies. We argued that most previous causal methods based on propensity score is 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. And we proposed a Data-Driven Variable Decomposition (D^{2}VD) algorithm to jointly optimize the variables decomposition and ATE estimation. Experimental results on synthetic data and real world data verify the practical usefulness of our model and the effectiveness of our D^{2}VD algorithm for ATE estimation in observational study.

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

References

Athey, S., and Imbens, G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

Bang, H., and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973.

Bloniarz, A.; Liu, H.; Zhang, C.-H.; Sekhon, J.; and Yu, B. 2016. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*.

Bottou, L.; Peters, J.; Quinonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14(1):3207–3260.

Brookhart, M. A.; Schneeweiss, S.; Rothman, K. J.; Glynn, R. J.; Avorn, J.; and Stürmer, T. 2006. Variable selection for propensity score models. *American journal of epidemiology* 163(12):1149–1156.

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

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. *arXiv preprint*.

Funk, M. J.; Westreich, D.; Wiesen, C.; Stürmer, T.; Brookhart, M. A.; and Davidian, M. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* 173(7):761–767.

Hernán, M. Á.; Brumback, B.; and Robins, J. M. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* 11(5):561–570.

Hernán, M. A.; Brumback, B. A.; and Robins, J. M. 2002. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine* 21(12):1689–1709.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81(396):945–960.

Kohavi, R., and Longbotham, R. 2011. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter* 12(2):31–35.

Lechner, M. 1999. Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics* 17(1):74–90.

Lewis, R., and Reiley, D. 2009. Retail advertising works! measuring the effects of advertising on sales via a controlled experiment on yahoo!

Lunceford, J. K., and Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23(19):2937–2960.

Parikh, N., and Boyd, S. 2013. Proximal algorithms. *Foundations and Trends in optimization* 1(3):123–231.

Pearl, J. 2009. *Causality: Models, Reasoning, and inderence*. Cambridge university press.

Reis, D.; Landeiro, V.; Culotta; and Aron. 2015. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 182–188.

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

Rosenbaum, P. R. 1987. Model-based direct adjustment. *Journal of the American Statistical Association* 82(398):387–394.

Sauer, B. C.; Brookhart, M. A.; Roy, J.; and VanderWeele, T. 2013. A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and drug safety* 22(11):1139–1145.

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

Sun, W.; Wang, P.; Yin, D.; Yang, J.; and Chang, Y. 2015. Causal inference via sparse additive models with application to online advertising. In *AAAI*.

VanderWeele, T. J., and Shpitser, I. 2011. A new criterion for confounder selection. *Biometrics* 67(4):1406–1413.