Proceedings of Machine Learning Research 129:811–822, 2020 ACML 2020

A New Representation Learning Method for Individual Treatment Effect Estimation: Split Covariate Representation Network

Qidong Liu LIUQIDONG@STU.XJTU.EDU.CN
Feng Tian FENGTIAN@MAIL.XJTU.EDU.CN
Weihua Ji JWH2017@STU.XJTU.EDU.CN
Qinghua Zheng QHZHENG@MAIL.XJTU.EDU.CN
Institute of Electronic and Information Science, Xi'an Jiaotong University, Xi'an, 710049, China

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Individual treatment effect (ITE) estimation is widely used in many essential fields, such as medical and education. But two problems, unknown counterfactual outcome and confounder, are the barriers for making a good ITE estimation. Although some representation learning methods based on potential outcome framework have been proposed to solve the problems, we find that most of previous works assume all features (also named covariate) of a unit are confounders. However, this assumption is not easy to become true, because instrument variables, adjustment variables and irrelevant variables can also be included in features. Therefore, this paper proposes a simple method to split covariates, and then a network, Split Covariate Representation Network (SCRNet), is mentioned, which is used to estimate ITE by different kinds of variables. Experiment results show that our method outperforms other state-of-arts methods on IHDP, a semi-synthetic dataset, and Jobs, a real-world dataset.

Keywords: Causal Inference, Individual Treatment Effect, Representation Learning

1. Introduction

Causal Inference gradually became a focus in machine learning community recently. As one of most important tasks in causal inference, ITE estimation plays a significant role in the fields of medical [Shalit (2019)], education [Zhao and Heffernan (2017)] and so on. Up to now, there are two basic models for counterfactual which are proved equivalent [Pearl (2009b)]: Structural Causal Model (SCM) [Pearl (2009a)] and Potential Outcome Framework (PO) [Rubin (2005)].

Except for unknown counterfactual outcome, confounder is the main barrier in ITE estimation. A confounder is a variable that not only affects treatment, but also affects outcome. Taking an assumed example in the field of education to illustrate how a confounder causes selection bias and disturbs treatment effect estimation, we want to know how the ranking, r, in college entrance examination affects undergraduates' grades, g, in a Chinese university. Obviously, *province*, as a variable where a student comes from, is a confounder. Because, r is a treatment, g is an outcome, and *province* affects both ranking and grades

due to different level of education in different province. However, students from high education quality province often rank lower than students from poor education quality province. Therefore, we could get the result that lower ranking leads to better grades. This error is resulted by imbalance distribution in treatment, which is also named selection bias.

For solving this problem, traditional representation methods based on PO framework regards selection bias as a domain adaptation problem. They learn a balance representation of confounders used to inference outcomes, which include factual outcome and counterfactual outcome. The most popular one of these methods is TARNet [Shalit et al. (2017)].

But we find most of previous works assumed that all features of a unit are confounders, however, such assumption is difficult to become true. We can see in Figure 1 that some covariates only affect treatment which are named instrument variables[Pearl (2009a)] and some covariates only affect outcome which are named adjustment variables[Pearl (2009a)]. Even, some covariates are irrelevant variables that have no effect on both treatment and outcome.

As for how to use these variables, we think that irrelevant variables should be removed at first when we estimate treatment effect. Because Pearl etc.[Pearl (2012)] have proved that conditioning on instrument variable will increase confounding bias. And in nonlinear model, conditioning on instrument even bring other new bias which is undesirable. In real world, causal model is often nonlinear. So, it is necessary to remove instrument variable when we estimate causal effect [Pearl (2012);Wooldridge (2016)]. Compared to confounders, we think adjustment variables are not necessary to be balanced. Previous works [Shalit et al. (2017);Johansson et al. (2016)] proposed that selection bias can be regarded as a case of domain adaptation and learning a balancing representation of confounders between treated and control group can solve this problem. However, adjustment variable has no effect on treatment, so it has no contribution to selection bias. We think it may bring some estimation bias when balancing the representation of adjustment blindly. Except for instrument variables and adjustment variables, some irrelevant variables are mixed in covariates. Though these variables have no effect on outcome prediction, we can decrease the collecting information about a unit by removing these variables when estimating causal effect. Causal inference is widely used in medication and education which both relate to privacy, such as family income and disease history. So, less covariates in ITE estimation can make collecting observational data easier.

To relax the assumption that all of covariates are confounders, we propose a simple method, which can divide covariates into confounders, instrument variables, adjustment variables and irrelevant variables. This method combines usual causal structure and conditional independence. Besides, we propose a new ITE estimation model, Split Covariate Representation Network (SCRNet), to make good use of split covariates. In a word, the contributions of this paper are:

- 1. Combining usual causal structure and conditional independence, this paper proposes a simple method to divide covariates into confounders, instrument variables, adjustment variables and irrelevant variables.

- 2. For removing the estimation bias caused by instrument variables, and for making good use of adjustment variables and confounders, a new ITE estimation model is proposed – Split Covariate Representation Network (SCRNet).

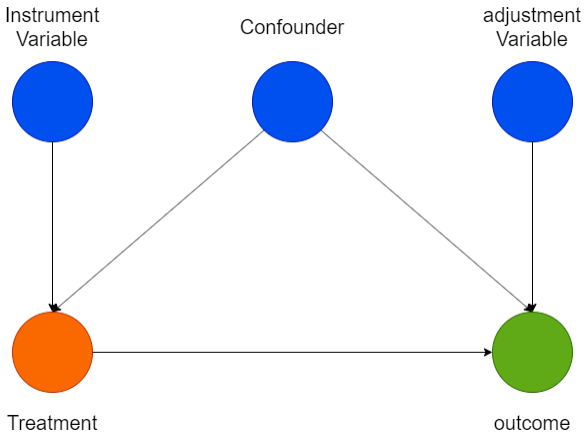

Figure 1: Covariates may include four kinds of variables: confounders, instrument variables, adjustment variables and irrelevant variables.

- 3. In experiment, two datasets are used to evaluate our method, and results show that our method outperforms other state-of-arts.

Some notations in this paper are given here. In this paper, we assume that the number of units is n, and each unit has features $x \in X$ and treatment t. The treatment is a binary variable as most of previous works setup. The units with $t = 1$ constitute treated group and the units with $t = 0$ constitute control group. Under PO framework, due to $t \in \{0, 1\}$, each unit has two potential outcomes, $y_i(t = 0)$ and $y_i(t = 1)$. One of potential outcomes is factual outcome, denoted as y^F, and the other is counterfactual outcome, denote as y^{CF}. After defining the potential outcome for ith unit, we give the definition of individual treatment effect (ITE):

$$ITE_i = y_i(t = 1) - y_i(t = 0)$$

Some assumptions are often made when using PO framework [Rubin (2005)]. And in this paper, we also follow these assumptions:

Assumption 1 (SUTVA). Every units should be independent of each other and has no interaction. Treatment has unique level.

Assumption 2 (Ignorability). Given Covariates X, potential outcomes should be independent of treatment assignment, i.e., $t \perp y(t = 1), y(t = 0)|X$

Assumption 3 (Positivity). For any set of covariates x, treatment is not decided, i.e., $P(t|X) > 0, \forall t, X$.

2. Related work

There are two main research directions in causal field – causal discovery [Kalisch and Buhlmann (2014);Guo et al. (2018)] and causal inference[Yao et al. (2020)]. The aim of causal discovery is to construct causal graph from data, such as observational data [Peter

et al. (2000)] and interventional data [Kocaoglu et al. (2017)]. Causal inference focus on treatment effect estimation. And PO framework is an important basis for this topic.

The methods based on PO framework can be divided into five categories. The first is re-weighting. These methods construct a supposed balance contribution by re-weighting each unit, such as DCB [Kuang et al. (2017a)]. The second category is stratification. The main idea of these methods is dividing units into different groups according to some standard so that the distribution is balance in each group. The third category is matching which is often used. In observational data, if a unit is treated, the nearest unit on x in control group is used as the counterfactual. For example, HSIC-NMM [Chang and Dy (2017)] and etc. belong to these matching methods. The forth category is tree-based method, which uses decision tree to learn how to predict counterfactual outcome. BART [Hill (2011)] and Causal Forest[Wager and Athey (2018)] are famous versions in this category.

And representation method is the last category which our method belongs to. Balancing Neural Network (BNN) [Johansson et al. (2016)] converted counterfactual prediction to domain adaptation firstly. Then, Treatment Agnostic Representation Network (TARNet) [Shalit et al. (2017)] and Counterfactual Regression (CFR) [Shalit et al. (2017)] used two-head network to improve performance. Recently, DragonNet [Shi et al. (2019)] added another head to predict treatment based on TARNet, and got a better result.

However, the methods mentioned above assumed that all confounders are observable. CEVAE [Louizos et al. (2017)] combined TARNet with auto-encoder, giving a method that can estimate treatment effect with latent confounders.

3. Method

For relaxing the assumption regarding all covariates as confounders, we use usual causal structure and conditional independence to split covariates. And then, this paper propose SCRNet with inputting confounders and adjustment variables to estimate treatment effect. So this section is divided into two parts to illustrate splitting covariates and SCRNet respectively.

3.1. Split Covariate

Each covariate of each unit can only be the cause of treatment and outcome, because covariates are inherent attributes of units. Therefore, each covariate probably belongs to one of confounder, instrument variable, adjustment variable and irrelevant variable. We assume that covariates X is set of variables, denoted as $X = \{x_1, x_2, \dots x_p\}$. And the sets of confounders, instrument variables and adjustment variables are denoted as X_c, X_i, X_a respectively.

Each item of X will be judged which kind of variable it belongs to. And when all items are judged, covariates are split. There are two steps in the process of judgement. The first step is to find out possible instrument variables and adjustment variables. As figure 2 shows that, instrument variable, treatment and outcome form a chain structure, while as figure 3 shows that treatment, outcome and adjustment variable form a collider structure. According to universal causal structure rule [Pearl (2009a)], start node is independent of end node in collider structure. So, if a covariate is not independent of treatment, the covariate cannot be an adjustment variable and we add x to X_i. However, start node interacts with

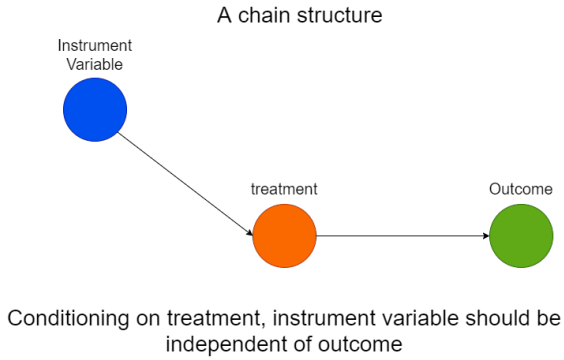

Figure 2: Instrument variable, treatment and outcome form a chain structure.

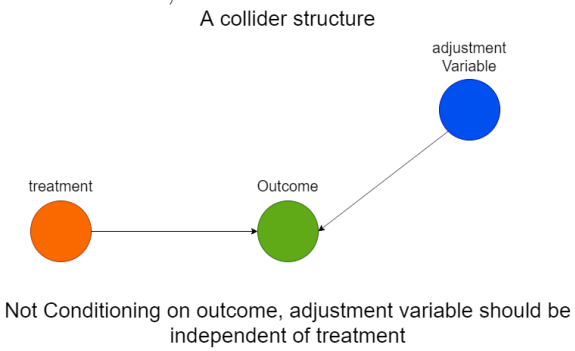

Figure 3: Treatment, outcome and adjustment variable form a collider structure.

end node under chain structure. If we block the middle node, this interaction will be cut down. And the method of block is conditional independence. Therefore, if we conditioning on treatment, instrument variable should be independent of outcome. Otherwise, there is other access between this covariate and outcome, which illustrates that this covariate has effect on outcome. So, for $x \in X_i$ which means that chain structure exits, if x is not independent of outcome conditioning on treatment, it may belong to adjustment variable and we add x to X_a. And for $x \notin X_i$, we only need judge whether x is independent of outcome.

The second step is to find out confounders. After the first step, variables in X_i and X_a are the cause of treatment and the cause of outcome respectively. For confounder is the cause of both treatment and outcome, the variables that belong to not only X_i but also X_a are confounders. So we add these covariates to X_c, which equals to $X_c = X_i \bigcap X_a$. And then removing covariates $x \in X_c$ from X_i and X_a, we will get the true instrument variables set and adjustment variables set. Irrelevant variables are the covariates that not belong to X_i, X_a and X_c. And split covariates are accomplished.

3.2. SCRNet

Recently, representation learning methods have succeeded in treatment effect estimation. The main idea of these methods is to convert the problem of selection bias to domain

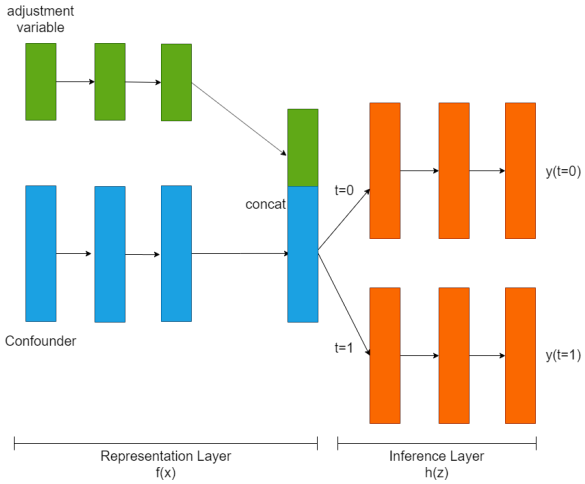

Figure 4: The structure of SCRNet.

adaptation. These methods learn a balance representation of confounders between treated group and control group. And almost all methods assume treatment as a binary variable, so a two-head network is proposed for predicting treated and control outcome respectively. With more precise outcome prediction, the performance of ITE estimation also becomes better.

SCRNet follows the basic structure of previous works. However, covariates are split instead of being confounders, so we have to alter the structure of network. Obviously, for irrelevant variables has no impact on outcome, we remove them directly. Although instrument variables can help estimate treatment effect when confounders are unobservable [Hartford et al. (2017)], they will bring bias when all confounders are observable[Pearl (2012);Wooldridge (2016)]. Therefore, the input of our model also excludes instrument variables. While adjustment variables have effect on outcome like confounders, they don't produce selection bias. So, we don't balance the representation of adjustment variables. And we simply concatenate their representation with confounders' representation. In conclusion, the structure of SCRNet is shown in figure 4.

As mentioned before, the aim of our model is to balance the representation of confounders and to reduce the error of outcome prediction. So, the loss of our model is:

$$\begin{aligned}\mathcal{L}(X_c, X_a, t) =& \frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot L(h(f_c(X_c^{(i)}), f_a(X_a^{(i)}), t_i), y_i) + \\ & \alpha \cdot Wass((f_c(X_c^{(i)}))_{i:t_i=0}, (f_c(X_c^{(i)}))_{i:t_i=1}) + \lambda \cdot ||h||_2\end{aligned} \tag{1}$$

$$\omega_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}, u = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{2}$$

Note that u is the proportion of treated units in the population. ω_i is named inverse probability of treatment weighting (IPTW) [Rosenbaum and Rubin (1983)], which is used to compensate the different size between treated and control group. And the calculation

of IPTW is shown in (2). In (1), $L(\cdot)$ is RMSE, which is used to reduce the prediction error.$Wass(\cdot,\cdot)$ is Wasserstein distance[Sriperumbudur et al. (2012)], which is used to measure the difference between two distributions. As a part of loss, it is used to balance the representation of confounders between treated group and control group. $||h||_2$ is the l2 regularization of parameters. α and λ are hyperparameters.

We used stochastic gradient descent to train our model by minimizing the loss in (1). During the process of training, the error is backpropagated to both inference layer and representation layer, which means that it is an end-to-end process. Two stage training like BNN [Johansson et al. (2016)] is not applied in our model.

4. Experiment

It is often difficult to evaluate causal inference algorithm, because counterfactual outcomes are often missed. With no counterfactual outcome, ground truth ITE is unknown. Therefore, most of previous works used synthetic or semi-synthetic datasets to conduct experiments. This paper applied a semi-synthetic dataset IHDP [Hill (2011)] and a real-world dataset Jobs [Shalit et al. (2017)] to evaluate our method.

Besides, we choose four baselines to compare with SCRNet. The first baseline is Balancing Neural Network (BNN) [Johansson et al. (2016)], which firstly adopted domain adaptation to causal inference. And the second is Treatment-Agnostic Representation Network (TARNet)[Shalit et al. (2017)], which applied two-head network to the inference layer. The third model is Counterfactual Regression (CFR)[Shalit et al. (2017)], which promoted performance by balancing representation of confounder. The last baseline is DragonNet [Shi et al. (2019)]. This algorithm add a head network to predict treatment, so the inference layer of DragonNet is a three-head network.

In experiment, we followed within-sample and out-of-sample setup as [Shalit et al. (2017)] did. The meaning of within-sample setup is to get accurate ITE when a unit has been treated or control, while the setup of out-of-sample is to make the best policy for a new unit.

4.1. Semi-Synthetic Dataset: IHDP

IHDP is a semi-synthetic dataset, which can be created by NPCI package [27]. And we generated 100 low sampled value (for accuracy of splitting covariates) IHDP replications for evaluation. In this dataset, treatment is specialist home visits and outcome is infants' future cognitive test scores. The number of covariates are 25, and these covariates are some information about infants and their mothers.

For there exists ground-truth ITE in IHDP, we can evaluate the precision of predicting ITE – Precision in Estimation of Heterogeneous Effects (PEHE) [Hill (2011)]. And the error of PEHE is shown below:

$$\varepsilon_{PEHE} = \frac{1}{N}\sum_{i=1}^{N}[(y_1^{(i)} - y_0^{(i)}) - (\hat{y}_1^{(i)} - \hat{y}_0^{(i)}))]^2 \tag{3}$$

Note that $\hat{y}_1^{(i)}$,$\hat{y}_0^{(i)}$ are the outcome of treated unit and control unit respectively. And ground-truth ITE is denoted as $y_1^{(i)} - y_0^{(i)}$.

Table 1: The results of experiment on IHDP.

Model	Within-sample	Out-of-sample
BNN	1.414	1.491
TARNet	0.669	0.716
CFR	0.654	0.706
DragonNet	0.753	0.786
SCRNet	**0.651**	**0.681**

The results of experiment on IHDP are shown in Table 1. Lower ε_{PEHE} represents better performance. Therefore, we can conclude that our method outperforms other state-of-arts on IHDP dataset. In addition, our method use less features of unit to get a better result, because we discard instrument variables and irrelevant variables, which is also an advantage of our method.

4.2. Real-world Dataset: Jobs

Jobs dataset is a real-world dataset, which consists of a randomized controlled trial (RCT) and an observational experiment. In this dataset, treatment is whether a person joins job training, and outcome represents whether a person has a job in 1978. Obviously, both treatment and outcome are binary variables. Covariates contain some information of job seekers, such as age, education and so on.

For there is no counterfactual outcome in this real-world dataset, ground-truth ITE is unknown. In [Shalit et al. (2017)], the paper proposed a standard, named policy risk, to evaluate causal inference algorithm. The definition of policy risk is:

$$R_{pol} = 1 - (E[Y_1|\pi_f(x) = 1] \cdot p(\pi_f = 1) + E[Y_0|\pi_f(x) = 0] \cdot p(\pi_f = 0)) \tag{4}$$

$$\pi_f(x) = 1, iff(x, 1) - f(x, 0) > \lambda. Otherwise, \pi_f(x) = 0 \tag{5}$$

In (4), $\pi_f(x)$ represents what policy we should make according to ITE estimator. And (5) gives how to calculate $\pi_f(x)$, with $\lambda = 0$ in this paper. The actual meaning of policy risk, denoted as R_{pol}, is average loss of decision implied by an ITE estimator.

Table 2: The results of experiment on Jobs.

Model	Within-sample	Out-of-sample
BNN	0.232	0.240
TARNet	0.228	0.234
CFR	0.223	**0.229**
DragonNet	0.226	0.235
SCRNet	**0.215**	**0.229**

We run experiments on jobs dataset 50 times and calculate the average R_{pol}. Table 2 shows the results of experiment on Jobs dataset. R_{pol} is lower, the performance of model

is better. According to the results, we can see that our model has a good performance compared with other state-of-arts.

4.3. Discussion

This paper focuses on splitting out variables of instrument, adjustment, irrelevant and confounders from covariates all at once. As we know, although one method [Yao et al. (2019)] based on representation learning has cared about the splitting issue, it used text-based sample description to split out instrument variables and removed them. But we don't introduce any other information to complete split task and realize a corresponding treatment effect estimator.And another work [Kuang et al. (2017b)] which split covariates into adjustment variables and confounders is aimed at average treatment effect (ATE) estimation.

Table 3: The number of different kinds of variable in two datasets.

Dataset	Total	Confounder	Adjustment	Instrument	Irrelevant
IHDP(average)	25	2.8	2.85	8	11.35
Jobs	7	6	1	0	0

As shown in Table 3, it is necessary to split out variables of instrument, adjustment, irrrelevant and confounder from covariates, which may improve the performance. We find that the performance of our method is better, in IHDP experiment, but is similar to others, in Jobs experiment. Compared with experimental results on Jobs, our method splits out more adjustment and instrument variables on IHDP dataset. If not splitting these variables, our method will be similar to CFR. Therefore, it is not difficult to understand why the result of our method is similar to CFR on Jobs dataset, for only one adjustment variable is split out. However, many variables of adjustment, instrument, and irrelevant are split on IHDP dataset, which illustrates that splitting covariates is necessary in ITE estimation. But the accuracy of independence test which is not high due to finite datasets limits our algorithm. In rare cases, we even have to split out confounders according to results of several confidence coefficient setup. So, more accurate independence testing would promote our algorithm furtherly. Moreover, we will study how different kinds of variables influence ITE estimation in the future.

5. Conclusion

In this paper, we have proposed a simple method for splitting covariates and a new ITE estimation model —SCRNet. Our method relaxes the assumption that all covariates of units are confounders, which most of previous works have made. And we evaluated our method on two datasets. The results of experiment indicate that our method outperforms other state-of-arts. In the future, we will focus on why our method can have a good performance with fewer features. And we think the combination with explainable machine learning is also promising.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2018YFB1004500), National Natural Science Foundation of China(61877048), Innovative Research Group of the National Natural Science Foundation of China(61721002), Innovation Research Team of Ministry of Education (IRT17R86), Project of China Knowledge Centre for Engineering Science and Technology, the consulting research project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", the Natural Science Basic Research Plan in Shaanxi Province of China under Grant No.2019JM458,2020JM-070, MoE-CMCC "Artificial Intelligence" Project No.MCM20190701.The Fundamental Research Funds for the Central Universities (sxzd012020003). The corresponding author is Feng Tian. And we thank all anonymous reviewers for their suggestions.

References

Yale Chang and Jennifer G Dy. Informative subspace learning for counterfactual inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1770–1776, 2017.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv: Artificial Intelligence*, 2018.

Jason Hartford, Greg Lewis, Kevin Leytonbrown, and Matt Taddy. Deep iv: a flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1414–1423, 2017.

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

Fredrik D Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

Markus Kalisch and Peter Buhlmann. Causal structure learning and inference: A selective review. *Quality Technology and Quantitative Management*, 11(1):3–21, 2014.

Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884, 2017.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 265–274, 2017a.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017b.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard S Zemel, and Max Welling. Causal effect inference with deep latent-variable models. pages 6446–6456, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009a.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009b.

Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv: Methodology*, 2012.

Spirtes Peter, N Glymour Clark, Scheines Richard, Heckerman David, Meek Christopher, Cooper Gregory, and Richardson Thomas. *Causation, prediction, and search*. MIT press, 2000.

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

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Uri Shalit. Can we learn individual-level treatment policies from clinical data? *Biostatistics*, 21(2):359–362, 2019.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *international conference on machine learning*, pages 3076–3085, 2017.

Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, 2019.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Jeffrey M Wooldridge. Should instrumental variables be used as matching variables. *Research in Economics*, 70(2):232–237, 2016.

Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. On the estimation of treatment effect with text covariates. In *roceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4106–4113, 2019.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *arXiv*, 2020.

Siyuan Zhao and Neil T Heffernan. Estimating individual treatment effect from educational studies with residual counterfactual networks. In *International Educational Data Mining Society*, 2017.