Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition

JUNKUNYUAN, Zhejiang University, China
ANPENG WU, Zhejiang University, China
KUN KUANG*, Zhejiang University, China
BO LI, Tsinghua University, China
RUNZE WU, NetEase Fuxi AI Lab, China
FEI WU, Zhejiang University, China
LANFEN LIN, Zhejiang University, China

Instrumental variables (IVs), sources of treatment randomization that are conditionally independent of the outcome, play an important role in causal inference with unobserved confounders. However, the existing IV-based counterfactual prediction methods need well-predefined IVs, while it's an art rather than science to find valid IVs in many real-world scenes. Moreover, the predefined hand-made IVs could be weak or erroneous by violating the conditions of valid IVs. These thorny facts hinder the application of the IV-based counterfactual prediction methods. In this paper, we propose a novel Automatic Instrumental Variable decomposition (AutoIV) algorithm to automatically generate representations serving the role of IVs from observed variables (IV candidates). Specifically, we let the learned IV representations satisfy the relevance condition with the treatment and exclusion condition with the outcome via mutual information maximization and minimization constraints, respectively. We also learn confounder representations by encouraging them to be relevant to both the treatment and the outcome. The IV and confounder representations compete for the information with their constraints in an adversarial game, which allows us to get valid IV representations for IV-based counterfactual prediction. Extensive experiments demonstrate that our method generates valid IV representations for accurate IV-based counterfactual prediction.

CCS Concepts: • **Computing methodologies** → **Causal reasoning and diagnostics**; *Machine learning*; *Statistical relational learning*.

Additional Key Words and Phrases: instrumental variable, counterfactual prediction, causal inference, representation learning, mutual information.

ACM Reference Format:
JunkunYuan, Anpeng Wu, Kun Kuang, Bo Li, Runze Wu, Fei Wu, and Lanfen Lin. 2018. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition. *J. ACM* 37, 4, Article 111 (August 2018), 20 pages. https://doi.org/10.1145/1122445.1122456

*Corresponding author.

Authors' addresses: JunkunYuan, yuanjk@zju.edu.cn, Zhejiang University, Zhejiang, China; Anpeng Wu, anpwu@zju.edu.cn, Zhejiang University, Zhejiang, China; Kun Kuang, kunkuang@zju.edu.cn, Zhejiang University, Zhejiang, China; Bo Li, libo@sem.tsinghua.edu.cn, Tsinghua University, Beijing, China; Runze Wu, wurunze1@corp.netease.com, NetEase Fuxi AI Lab, Zhejiang, China; Fei Wu, wufei@cs.zju.edu.cn, Zhejiang University, Zhejiang, China; Lanfen Lin, llf@zju.edu.cn, Zhejiang University, Zhejiang, China.

xxxx-xxxx/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As a representative task in machine learning [12, 13, 22, 35], supervised learning [9, 38] explores correlations between variables from rich data for prediction. However, in many real applications, a decision-maker always wants to judge the counterfactual impact of treatment (policy) changes on the outcome that can not be found in the data. For example, an airline wants to estimate the effect of prices (i.e. treatment) on customers' purchase tendency (i.e. outcome) [18]. We may observe that examples with high prices are often associated with high sales in data sampled during holidays, which may fool the direct supervised learning approaches to predict that increasing prices would also lead to high sales at other times. In this case, we can add the observable confounders (i.e., holidays, which cause the changes in both the prices and the sales) into training data to correct the model. Nevertheless, if there exist unobserved confounders (e.g., conferences, which are also common causes of the prices and the sales but are unknown to the decision-maker), the typical supervised learning model would still head in the wrong direction.

Instrumental Variables (IVs) [43] are exogenous variables that are correlated to the treatment but do not directly affect the outcome, which provides an alternative approach for counterfactual prediction even with the unobserved confounders. Existing IV-based counterfactual prediction methods mainly adopt a two-stage procedure, which first builds a model to estimate the treatment based on the IVs, and then predicts the outcome with the estimated treatment. Two-stage least squares (2SLS) [2] is a well-known method that employs the two-stage procedure with linear models and obtains homogeneous treatment effects. Recent IV-based counterfactual prediction works [3, 18, 30, 31, 37] mainly focus on generalizing previous approaches on high-dimensional and non-linear data. These methods achieve great counterfactual prediction performance, however, they rely heavily on well-predefined IVs. In many real-world applications, we can hardly have enough prior knowledge to identify the valid IVs [28] (i.e., the variables that satisfy the relevance, the exclusion, and the unconfounded instrument conditions, see Sec. 3 for details). Moreover, the predefined hand-made IVs could be weak or erroneous by violating some of the conditions of the valid IVs. Therefore, it's highly demanding to develop a data-driven approach to automatically obtain valid IVs (or IV representations) for the downstream IV-based counterfactual prediction methods.

In many real applications, although there are always a large number of observed variables, few of them satisfy the conditions of the valid IVs. Since finding the valid IVs is difficult, instead, there are growing works that focus on synthesizing valid summary IVs with IV candidates [28] (some of them might be invalid IVs, i.e., do not strictly satisfy the conditions of the valid IVs). Mendelian Randomization (MR) [4] is a popular approach that utilizes genetic markers as the IVs to perform causal inference [46] among clinical factors. Unweighted/Weighted Allele Scores (UAS/WAS) [6, 7, 11] that weigh each IV candidate equally or based on the correlation between them and the treatment are representative methods in MR. However, they need all the IV candidates to be both valid and independent conditional on the summary IVs. Hartford et al. [17] apply an ensemble method to select valid IV set with asymptotical validity. But it not only relies on the independence and modal validity of IV candidates but also needs high computation costs by running the downstream IV-based methods with every IV candidate for valid set selection. Kuang et al. [28] present to model a summary IV as a latent variable and estimate it by utilizing recent advances in weak supervision that is based on statistical dependencies among the IV candidates. However, this method is confined to the binary variable setting, limiting its use in many real-world applications.

Inspired by the recent works [19, 44, 47] on causal disentangled representation learning, we argue that although invalid IV candidates do not satisfy the conditions of the valid IVs strictly, one might decompose and utilize a part of their information to generate IV representations. Therefore, in this

paper, we propose a novel Automatic Instrumental Variable decomposition (AutoIV) algorithm to automatically generate representations serving the role of IVs for counterfactual prediction with fewer constraints for the IV candidates. Specifically, we first generate the IV representations from the IV candidates and make them satisfy the relevance condition with the treatment and the exclusion condition with the outcome via mutual information maximization and minimization constraints, respectively. We also generate confounder representations by encouraging them to be relevant to both the treatment and the outcome. The IV and the confounder representations compete for the corresponding information with their constraints in an adversarial game, which allows us to obtain valid IV representations for counterfactual prediction with the downstream IV-based methods.

In summary, the main contributions of this paper are:

- We study the problem of IV-based counterfactual prediction under a more practical setting, i.e., no valid IVs are available for learning, which is beyond the capability of the previous IV-based methods.
- We propose a novel Automatic Instrumental Variable decomposition (AutoIV) algorithm to automatically generate IV representations that satisfy the conditions of the valid IVs from the IV candidates. It adopts mutual information constraints to control representation learning process via an adversarial game.
- Extensive experiments show that the proposed method generates valid IV representations for accurate counterfactual prediction, which is even comparable to directly use the true valid IVs.

The rest of the paper is organized as follows. In Sec. 2, some related works about IV-based counterfactual prediction, IV synthesis, and causal representation learning are introduced. In Sec. 3, the definition of the valid IVs and some related IV-based methods are stated. In Sec. 4, our automatic instrumental variable decomposition algorithm is introduced. In Sec. 5, the results of the experiments on low-dimensional and high-dimensional are reported. We discuss the investigation with a future research outlook in Sec. 6.

2 RELATED WORK

In this section, we briefly review the related works of IV-based counterfactual prediction, IV synthesis, and causal representation learning in recent years.

2.1 IV-based Counterfactual Prediction

Two-stage least squares (2SLS) [2] is a representative method for IV-based counterfactual prediction with linear models in causal inference researches [1, 24, 26, 46, 48]. Many recent IV-based counterfactual prediction methods extend 2SLS to non-linear and high-dimensional settings. One research direction is the generalized method of moments (GMM) [16], which uses moment conditions to estimate model parameters. A recent trend is to combine GMM with machine learning, like selecting moment conditions via adversarial training [30] and variational reformulation of GMM with deep neural networks [3]. Another direction is based on kernel approaches, such as a single-stage kernel approach [31] and a novel method with consistency guarantees [37]. DeepIV [18] is a recent remarkable study that fits a mixture density network for the treatment and trains an outcome prediction model with the estimated conditional treatment distribution. All of the above methods need predefined IVs, and their performance relies on the validity of the given IVs. However, identifying and obtaining valid IVs may be thorny because their validity conditions are strict.

2.2 IV Synthesis

There are growing works [4, 5, 5, 15, 17, 21, 28, 40] that propose to synthesize a valid summary IV by using the given observed variables (IV candidates) in recent years. Among them, some works [4, 5] are based on the independence condition of IV candidates, which is a strong restrictive property [28]. Some approaches [5, 15, 21, 40] perform reliable estimation only when most of the IV candidates are valid, which is also a strong condition. Hartford et al. [17] adopt ensemble methods based on the modal validity of the IV candidates, however, it needs expensive computation cost to select the valid IV set. Unweighted/Weighted Allele Scores (UAS/WAS) [6, 7, 11] weigh each IV candidate equally or based on the correlation between them and the treatment. Kuang et al. [28] generalize the allele scores method [6, 7, 11], which builds a summary IV and estimates it with advanced methods from weak supervision and structure learning. However, it only applies to the binary variable setting. These previous IV synthesis methods rely on some strong conditions for the IV candidates and may not be practical in many real scenes, while we present an automatic IV representation learning algorithm that only needs mild assumptions in this paper. Take the airline case as an example. When we are looking for valid IVs, e.g., fuel costs, from the IV candidates, we do not need to assume that they are valid, modal validity, or binary, but only need them to be correlated with the treatment, i.e., price, and be independent of the unobserved confounders, i.e., conferences.

2.3 Causal Representation Learning

Recently, causal representation learning [19, 20, 25, 26, 36, 44, 47] has attracted lots of attention in many applications [23, 27, 33, 39, 45, 49]. Among these works, Yao et al. [47] propose to reduce prediction bias by filtering out the nearly IVs. Some works [19, 44] decompose the IV, confounder, and adjustment representations by encouraging or limiting the correlations between variables. However, these works are limited to the binary treatment setting. Moreover, they neither give empirical results to show the effectiveness of the learned IV representations nor make use of the decomposed IV representations for counterfactual prediction. In contrast, we present a data-driven IV representation learning algorithm and show its effectiveness by applying the learned representations to the downstream IV-based methods for accurate counterfactual prediction.

3 PRELIMINARY

By following previous works [3, 37], we assume the relationship between treatment variable X and outcome variable Y in data generating process is

$$Y = g(X) + e, \tag{1}$$

where $g(\cdot)$ is an unknown causal response function which is potentially non-linear and continuous, and e is the error term that contains unobserved latent factors (i.e. unmeasured confounders) which affect both X and Y. Here, we assume the error term e is with zero expectation and finite variance (i.e., $\mathbb{E}[e] = 0$ and $\mathbb{E}[e^2] < \infty$). e contains unobserved factors that affect X, thus e would be correlated with X, i.e. $\mathbb{E}[e|X] \neq 0$, which makes X an endogenous variable and leads to $g(X) \neq \mathbb{E}[Y|X]$. Thus, it is infeasible to estimate the causal relationship $g(\cdot)$ between X and Y via directly estimating $\mathbb{E}[Y|X]$ from data distribution $P(X, Y)$ because of the confounding effect caused by the unobserved error e. The instrumental variables (IVs) are introduced to solve the endogenous treatment problem as we introduced previously. Valid IVs (denoted by Z) should satisfy the following conditions [3, 18, 37]:

- **Relevance.** Z is related to X, i.e., $\mathbb{P}(X|Z) \neq \mathbb{P}(X)$;
- **Exclusion.** Z is related to Y only through X and e, i.e., $\mathbb{P}(Y|Z, X, e) = \mathbb{P}(Y|X, e)$;
- **Unconfounded Instrument.** Z should be unconfounded, i.e., $\mathbb{E}[e|Z] = \mathbb{E}[e]$.

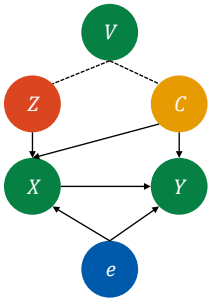

Fig. 1. The proposed AutoIV framework. Variables V, X, and Y are corresponding to the observed variables, treatment, and outcome, respectively. Variables e are unobserved confounders that are related to both X and Y. AutoIV decomposes representations of instrumental variables Z and confounders C from the observed variables V automatically, then use the learned representations for IV-based counterfactual prediction.

The goal of IV-based counterfactual prediction is to obtain a counterfactual estimation function $\hat{g}$ that is close to the true response function g. Moreover, if there exists exogenous variable C (i.e., $\mathbb{P}(e|C) = \mathbb{P}(e)$), we can make use of it for more accurate estimation, i.e. $X = (X', C)$ and $Z = (Z', C)$, where X' and Z' are the true treatment variable and instrumental variable, respectively. Note that we will also learn confounder representations in our algorithm, which are used as the exogenous variables C in the IV-based counterfactual prediction process.

Previous IV-based counterfactual prediction approaches assume that they have access to the true valid IVs Z which strictly satisfy the above conditions. Then, we could identify the causal response function $g(\cdot)$ based on

$$\mathbb{E}[Y|Z] = \mathbb{E}[g(X)|Z] = \int g(X) d\mathbb{P}(X|Z). \tag{2}$$

That is, one may first learn $\mathbb{P}(X|Z)$, then use it to estimate $g(\cdot)$. For example, standard two-stage least squares (2SLS) method [2] first learns $\mathbb{E}[\phi(X)|Z]$ with linear basis $\phi(\cdot)$, then fits Y by least-squares regression with the coefficient $\hat{\phi}(\cdot)$ that estimated in the first stage. Some non-parametric works [10, 32] extend the model basis to more complicated mapping function or regularization, e.g. polynomial basis. DeepIV [18] is proposed to apply deep neural networks in the two-stage procedure. It fits a mixture density network $F_\phi(X|Z)$ in the first stage and regresses Y by sampling from the estimated mixture Gaussian distributions of X. KernelIV [37] is a recent kernel approach that maps Z, X, and Y to reproducing kernel Hilbert spaces and perform the two-stage procedure in that space. DeepGMM [3] extends the existing GMM methods in the high-dimensional treatment and IVs setting, which is based on a novel variational reformulation of the optimally-weighted GMM.

The above existing IV-based counterfactual prediction methods need well-predefined valid IVs. However, it is an art rather than science to find suitable IVs in real applications. Even worse, the predefined hand-made IVs could be weak or erroneous by violating the conditions. Without the valid IVs, the counterfactual prediction performance of these downstream IV-based methods cannot be guaranteed.

In this paper, we aim to automatically learn valid IV representations that can be applied to the downstream IV-based methods for accurate counterfactual prediction. The validity of the learned IV representation determines the accuracy of the downstream counterfactual prediction task.

4 METHOD

In this section, we propose a novel Automatic Instrumental Variable decomposition (AutoIV) algorithm to generate decomposed IV and confounder representations from the observed variables. The proposed framework of AutoIV is shown in Fig. 1. The green part represents all the available variables, including observed variables V, treatment variables X, and outcome variables Y. E denotes unobserved confounders that are related to both X and Y. The observed variables V are correlated with X and also might be associated with Y. Similar to the general setting in recent IV analysis works [3, 18, 37], observed variables V are assumed exogenous, i.e. $\mathbb{P}[E|V] = \mathbb{P}[E]$. Therefore, the decomposed representations of instrumental variables Z and confounders C are also exogenous, which satisfies the unconfounded IV condition. Suppose that we have data $\mathcal{D} = \{(\boldsymbol{v}_i, \boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, our goal is to learn the representations of Z and C from the observed variables V based on their relationships to X and Y with data $\mathcal{D}$. Then, we use the learned representations for counterfactual prediction with the downstream IV-based methods introduced in Sec. 3. The validity of the learned representations determines the accuracy of the IV-based counterfactual prediction.

We first use neural networks to model the representations for Z and C as $\phi^Z(\cdot)$ and $\phi^C(\cdot)$ with parameters θ_{ϕ^Z} and θ_{ϕ^C}, respectively. The observed variables V are used as inputs of the representation networks. We control the information that flows into $\phi^Z(\cdot)$ to be related to X and conditionally independent of Y, which is based on the relevance and exclusion conditions, respectively. We then let $\phi^C(\cdot)$ be related to both X and Y. These two representation networks compete for the corresponding information with their constraints in an adversarial game. A general two-stage counterfactual prediction loss is then employed to further calibrate the learned representations.

Let A and B be two random variables that are correlated with each other. We have examples $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ sampled from the distributions of A and B, respectively. We encourage (or discourage) the relevance between A and B by maximizing (or minimizing) the mutual information between them. However, only the samples $\{(\boldsymbol{a}_i, \boldsymbol{b}_i)\}_{i=1}^N$ are available in our task, but what mutual information estimation need is data distributions. Inspired by recent works on contrastive learning and sample-based mutual information estimation [8, 34], we first learn a variational distribution $q(B|A)$ to approximate $\mathbb{P}(B|A)$. We let positive sample pair to be the sample pair with the same index (i.e. (a_i, b_i)), and let negative sample pair to be the sample pair with the different index $(a_i, b_j)_{i\neq j}$. As we already have the variational approximation $q(B|A)$, we can increase (or decrease) the relevance between A and B by maximizing (or minimizing) the differences between the variational approximation of the positive sample pair (i.e. $q(b_i|a_i)$) and that of the negative sample pair (i.e. $q(b_j|a_i)$). It can intuitively be interpreted that mutual information maximization task is achieved when there exists distinct differences between the relevance of a_i to its corresponding b_i and the relevance of a_i to b_j (where $i \neq j$). Meanwhile, mutual information minimization is to reduce that differences. Although there is deviation between $q(B|A)$ and $\mathbb{P}(B|A)$, the estimated mutual information is still excellent with great variational approximation [8].

4.1 Learning IV Representations

We aim to learn the IV representations that satisfy the conditions of the valid IVs (see Sec. 3), i.e., relevance, exclusion, and unconfounded instrument. Since we have already assumed the exogeneity of the observed variables V by following previous works [3, 18, 37], and the learned representations always satisfy the unconfounded instrument condition, we only need to make the learned IV representations satisfy the relevance condition with the treatment and the exclusion condition with the outcome.

Learning relevance. The relevance condition, i.e., $\mathbb{P}(X|Z) \neq \mathbb{P}(X)$, require IV representations $\phi^Z(V)$ to be correlated with the treatment X. Therefore, we encourage the information of V that is related to X to enter the IV representations $\phi^Z(V)$. We first use variational distribution $q_{\theta_{ZX}}(X|\phi^Z(V))$ with neural network parameters θ_{ZX} to approximate the true conditional distribution $\mathbb{P}(X|\phi^Z(V))$. The log-likelihood loss function of variational approximation $q_{\theta_{ZX}}(X|\phi^Z(V))$ with N samples is given as:

$$\mathcal{L}_{ZX}^{LLD} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{ZX}}(\boldsymbol{x}_i|\phi^Z(\boldsymbol{v}_i)). \tag{3}$$

We minimize Eq. (3) to get optimal variational approximation $q_{\hat{\theta}_{ZX}}(X|\phi^Z(V))$ with parameters $\hat{\theta}_{ZX}$. To increase the relevance between the IV representations and the treatment, we maximize the mutual information between them with

$$\mathcal{L}_{ZX}^{MI} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\log q_{\theta_{ZX}}(\boldsymbol{x}_i|\phi^Z(\boldsymbol{v}_i)) - \log q_{\theta_{ZX}}(\boldsymbol{x}_j|\phi^Z(\boldsymbol{v}_i))), \tag{4}$$

where $\log q_{\theta_{ZX}}(\boldsymbol{x}_i|\phi^Z(\boldsymbol{v}_i))$ represents the conditional log-likelihood of positive sample pair $(\phi^Z(\boldsymbol{v}_i), \boldsymbol{x}_i)$ and $q_{\theta_{ZX}}(\boldsymbol{x}_j|\phi^Z(\boldsymbol{v}_i)_{i\neq j}$ represents the negative sample pair $(\phi^Z(\boldsymbol{v}_i), \boldsymbol{x}_j)_{i\neq j}$. We minimize Eq.(4) to optimize the IV representations $\phi^Z(V)$ for relevance condition via maximizing differences between the positive and negative sample pairs.

Learning exclusion. The exclusion condition requires IV representations to be related to the outcome Y only through the treatment X and unobserved error e, i.e. $\mathbb{P}(Y|Z,X,e) = \mathbb{P}(Y|X,e)$. Since e is unobserved, we employ a more strict condition instead, i.e., $Z \perp\!\!\!\perp Y|X$. Therefore, we minimize mutual information between Z and Y conditional on X. Similarly, we first use variational distribution $q_{\theta_{ZY}}(Y|\phi^Z(V))$ with parameters θ_{ZY} to approximate the true conditional distribution $\mathbb{P}(Y|\phi^Z(V))$. The log-likelihood loss function for $q_{\theta_{ZY}}(Y|\phi^Z(V))$ is given as

$$\mathcal{L}_{ZY}^{LLD} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{ZY}}(\boldsymbol{y}_i|\phi^Z(\boldsymbol{v}_i)). \tag{5}$$

The optimal variational approximation $q_{\hat{\theta}_{ZY}}(\boldsymbol{y_i}|\phi^Z(\boldsymbol{v_i}))$ is achieved with parameters $\hat{\theta}_{ZY}$ by minimizing Eq. (5). The IV representations $\phi^Z(V)$ should be independent of the outcome Y given the treatment X, we achieve it by minimizing the mutual information between them. Since the treatments X are continuous random variables, we consider the constraints of conditional independence with smooth weight w_{ij}, and the loss function for mutual information minimization between IV representations $\phi^Z(V)$ and the outcome Y is given as:

$$\mathcal{L}_{ZY}^{MI} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\omega_{ij}\cdot(\log q_{\theta_{ZY}}(\boldsymbol{y}_i|\phi^Z(\boldsymbol{v}_i)) - \log q_{\theta_{ZY}}(\boldsymbol{y}_j|\phi^Z(\boldsymbol{v}_i)))). \tag{6}$$

Different from mutual information maximization in learning relevance, we let the positive $((\phi^Z(\boldsymbol{v}_i), \boldsymbol{y}_i))$ and negative $(\phi^Z(\boldsymbol{v}_i), \boldsymbol{y}_j)$ sample pairs have close log-likelihood expectation to make the IV representations $\phi^Z(V)$ and the outcome Y conditional independent. ω_{ij} is the weight of each pair of positive and negative samples, and we determine it by the discrepancy between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in RBF kernel:

$$\omega_{ij} = \text{softmax}(e^{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}}), \quad i,j = 1,2,...,N, \tag{7}$$

where σ is a hyperparameter, we use 0.5 for it in our experiments. The weight of positive and negative sample pairs increases when their treatments X have closer distance. In other words, we

would like to pay attention to the pairs which have close X values for our conditional independent constraints.

4.2 Learning Confounder Representations

We also decompose and learn the representations of confounders that are correlated to both the treatment and outcome. They are used as exogenous variables C for counterfactual prediction (see Sec. 3). We let the generated confounder representations, i.e. $\phi^C(V)$, are both correlated to the treatment X and outcome Y variables. With the similar procedure in learning IV representations, we first use variational distribution $q_{\theta_{CX}}(X|\phi^C(V))$ to approximate conditional distribution $\mathbb{P}(X|\phi^C(V))$, and the corresponding log-likelihood loss function is given as:

$$\mathcal{L}_{CX}^{LLD} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{CX}}(\boldsymbol{x}_i|\phi^C(\boldsymbol{v}_i)), \tag{8}$$

Optimal approximation $q_{\hat{\theta}_{CX}}(X|\phi^C(V))$ with parameter $\hat{\theta}_{CX}$ is obtained by minimizing (8). We then minimize the loss function of mutual information maximization between confounder representations $\phi^C(V)$ and the treatment X:

$$\mathcal{L}_{CX}^{MI} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\log q_{\theta_{CX}}(\boldsymbol{x}_i|\phi^C(\boldsymbol{v}_i)) - \log q_{\theta_{CX}}(\boldsymbol{x}_j|\phi^C(\boldsymbol{v}_i))). \tag{9}$$

The pairs of positive sample $(\phi^C(\boldsymbol{v}_i), \boldsymbol{x}_i)$ and negative sample $(\phi^C(\boldsymbol{v}_i), \boldsymbol{x}_j)$ are used to increase the relevance between C and X. Also, the variational distribution $q_{\theta_{CY}}(Y|\phi^C(V)$ for conditional distribution $\mathbb{P}(Y|\phi^C(V))$ and its mutual information maximization loss function is given as:

$$\mathcal{L}_{CY}^{LLD} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{CY}}(\boldsymbol{y}_i|\phi^C(\boldsymbol{v}_i)), \tag{10}$$

$$\mathcal{L}_{CY}^{MI} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\log q_{\theta_{CY}}(\boldsymbol{y}_i|\phi^C(\boldsymbol{v}_i)) - \log q_{\theta_{CY}}(\boldsymbol{y}_j|\phi^C(\boldsymbol{v}_i))). \tag{11}$$

We minimize Eq. (10) to get optimal variational approximation $q_{\hat{\theta}_{CY}}(Y|\phi^C(V))$ with parameter $\hat{\theta}_{CY}$, and minimize Eq. (11) to encourage the confounder representations $\phi^C(V)$ and the outcome Y to be relevant.

Since conditional on the confounders that contain IV information would introduce bias in causal inference [42], also, if the information of confounders (i.e. the variables correlated to Y) is embedded in the IV representations would influence the exclusion condition. Therefore, we minimize mutual information between the IV representations $\phi^Z(V)$ and confounder representations $\phi^C(V)$ to regularize the learned information in the generated representations. The variational distribution $q_{\theta_{ZX}}(\phi^C(V)|\phi^Z(V)$ for conditional distribution $\mathbb{P}(\phi^C(V)|\phi^Z(V))$ and the mutual information minimization loss function are given as:

$$\mathcal{L}_{ZC}^{LLD} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{ZC}}(\phi^C(\boldsymbol{v}_i)|\phi^Z(\boldsymbol{v}_i)), \tag{12}$$

$$\mathcal{L}_{ZC}^{MI} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\log q_{\theta_{ZC}}(\phi^C(\boldsymbol{v}_i)|\phi^Z(\boldsymbol{v}_i)) - \log q_{\theta_{CY}}(\phi^C(\boldsymbol{v}_j)|\phi^Z(\boldsymbol{v}_i))). \tag{13}$$

We minimize Eq. (12) to learn accurate variational approximation $q_{\theta_{ZX}}(\phi^C(V)|\phi^Z(V)$ for the conditional distribution $\mathbb{P}(\phi^C(V)|\phi^Z(V))$, and use the variational approximation to regularize the IV and confounder representations via minimizing Eq. (13).

In the above procedure with mutual information constraints, the IV representations $\phi^Z(V)$ attempt to extract information that is correlated to the treatment X and conditional independent to the outcome Y, while the confounder representations $\phi^C(V)$ are encouraged to be correlated to both X and Y. We also employ a regularization term to encourage the information to enters one of the extracted representations. Therefore, the two representation networks compete for the corresponding information with their constraints in an adversarial game, which allows us to get valid IV and confounder representations. We then introduce the general IV-based counterfactual prediction procedure to further improve the learned representations in the following.

4.3 Representation Calibration

We combine mutual information-based representation learning with a general two-stage counterfactual prediction procedure to further calibrate the learned representations. More concretely, we first regress X on IV and confounder representations, i.e., $\phi^Z(V)$ and $\phi^C(V)$,

$$\mathcal{L}_X = \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{x}_i, f^X(\phi^Z(\boldsymbol{v}_i), \phi^C(\boldsymbol{v}_i))), \tag{14}$$

where f^X is the first-stage (treatment) regression network with parameter θ_{f^X}, and $l(\cdot,\cdot)$ measures square error in our experiments. We then use the estimated treatment $\hat{X}$ (in the first stage) to regress the outcome Y in the second stage:

$$\mathcal{L}_Y = \frac{1}{N}\sum_{i=1}^{N} l(y_i, f^Y(\phi^C(v_i), f^{emb}(f^X(\phi^Z(v_i), \phi^C(v_i))))), \tag{15}$$

where f^{emb} is an embedding network with parameter $\theta_{f^{emb}}$ for expanding the dimension of $\hat{X}$, f^Y is the second-stage (outcome) regression network with parameter θ_{f^Y}. $\mathcal{L}_{X_r}$ and $\mathcal{L}_{Y_r}$ are minimized to optimize the paramters of representation, treatment, embedding, and outcome networks to further improve the decomposed representations.

Note that we assume that the candidate IVs are independent of the unobserved confounders. Based on our regularization term, the decomposed IV representations meet the relevance and exclusion assumptions. Besides, effect homogeneity and monotonicity assumption are often used in the analysis of instrumental variables. Based on the structural equation model, our algorithm model a homogeneity IV to estimate the accurate structural function of the treatment on the outcome [14, 41, 43].

4.4 Model Optimization

As we minimize Eq. (3), (5), (8), (10), and (12) to optimize the parameters θ_{ZX}, θ_{ZY}, θ_{CX}, θ_{CY}, and θ_{ZC}, respectively, each variational distribution approximates the corresponding conditional distribution. We simplify the expression by combining all the variational approximation loss as

$$\mathcal{L}^{LLD} = \mathcal{L}^{LLD}_{ZX} + \mathcal{L}^{LLD}_{ZY} + \mathcal{L}^{LLD}_{CX} + \mathcal{L}^{LLD}_{CY} + \mathcal{L}^{LLD}_{ZC}. \tag{16}$$

Notice that each loss term in Eq. (16) optimize the corresponding parameters and will not interact with each other. We then combine all the mutual information contraints loss functions of Eq. (4), (6), (9), (11), and (13) as

$$\mathcal{L}^{MI} = \mathcal{L}^{MI}_{ZX} + \mathcal{L}^{MI}_{ZY} + \alpha(\mathcal{L}^{MI}_{CX} + \mathcal{L}^{MI}_{CY}) + \eta\mathcal{L}^{MI}_{ZC}, \tag{17}$$

where α and η are hyper-parameters tuned on a held-out validation set. Eq. (17) is minimized to optimize the representation networks $\phi^Z(\cdot)$ and $\phi^C(\cdot)$ with parameters θ_{ϕ^Z} and θ_{ϕ^C}. Eq. (14) is minimized to optimize parameters of the representation and treatment networks (i.e., θ_{ϕ^Z}, θ_{ϕ^C}, and θ_{f^X}), and Eq. (15) is minimized to optimize the parameters of the representation, embedding,

Algorithm 1 AutoIV: Automatic IV Decomposition

Input: Training set $\mathcal{T} = (\boldsymbol{v}_i, \boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^{N_T}$; variational distribution parameters θ_{ZX}, θ_{ZY}, θ_{CX}, θ_{CY}, and θ_{ZC}; IV and confounder representation networks $\phi^Z(\cdot;\theta_{\phi^Z})$ and $\phi^C(\cdot;\theta_{\phi^C})$, respectively; treatment regression, embedding, and outcome regression networks $f^X(\cdot;\theta_{f^X})$, $f^{emb}(\cdot;\theta_{f^{emb}})$, and $f^Y(\cdot;\theta_{f^Y})$, respectively; hyperparameters α and η; training epochs M; batchsize B.

Output: Well-trained $\phi^Z(\cdot;\hat{\theta}_{\phi^Z})$ and $\phi^C(\cdot;\hat{\theta}_{\phi^C})$

1: Initialize Adam optimizer and all the parameters;
2: **for** *epoch* = 1 to M **do**
3: Randomly sample B examples from $\mathcal{T}$;
4: Update variational distribution parameters θ_{ZX}, θ_{ZY}, θ_{CX}, θ_{CY}, θ_{ZC} by minimizing $\mathcal{L}^{LLD}$ as Eq. (16);
5: Update representation networks parameters θ_{ϕ^Z} and θ_{ϕ^C} by minimizing $\mathcal{L}^{MI}$ as Eq. (17);
6: Update representation and treatment regression network parameters θ_{ϕ^Z}, θ_{ϕ^C}, θ_{f^X} by minimizing $\mathcal{L}_X$ as Eq. (14);
7: Update representation, embedding, and outcome regression network parameters θ_{ϕ^Z}, θ_{ϕ^C}, $\theta_{f_{emb}}$, θ_{f^Y} by minimizing $\mathcal{L}_Y$ as Eq. (15).
8: **end for**

and outcome networks (i.e., θ_{ϕ^Z}, θ_{ϕ^C}, $\theta_{f^{emb}}$, and θ_{f^Y}). We optimize Eq. (16), (17), (14), and (15) for the corresponding parameters alternately to get optimal decomposed representations of IVs and confounders. Finallym, we use the generated representations for counterfactual prediction with downstream IV-based methods to testify the validity of the learned representations. The whole optimization procedure of our AutoIV algorithm is stated in Algorithm 1.

5 EXPERIMENTS

In this section, we show the empirical evaluation of applying AutoIV to different downstream IV-based methods for counterfactual prediction. The validity of the learned IV representations determines the accuracy of counterfactual prediction of the downstream methods. We implement the experiments with Python on a device with CPU Intel Xeon Gold 6254, GPU Nvidia RTX 2080TI, and memory 64MB.

We list the representative IV-based methods introduced previously and used in our experiments in the following.

1. **DirectNN**: directly regress the outcome on the treatment with neural networks. It does not use any information of the IVs, and can be considered as a general supervised learning.
2. **2SLS (van)**: vanilla two-stage least squares with linear models.
3. **2SLS (poly)**: two-stage least squares with polynomial basis and ridge regularization.
4. **2SLS (NN)**: two-stage regression with neural networks structure.
5. **DeepIV** [18]: fit the treatment with the IVs via optimizing a mixture density networks in the first stage, and then fit the outcome by sampling from the mixture density networks. We use its original implementation [1].
6. **KernelIV**[37]: a recent kernel method that performs two-stage procedure in reproduce kernel Hilbert spaces. We implement it with Python by referring its original MATLAB version [2]. The results of ours and original MATLAB version are consistent.

[1] https://github.com/jhartford/DeepIV
[2] https://github.com/r4hu1-5in9h/KernelIV

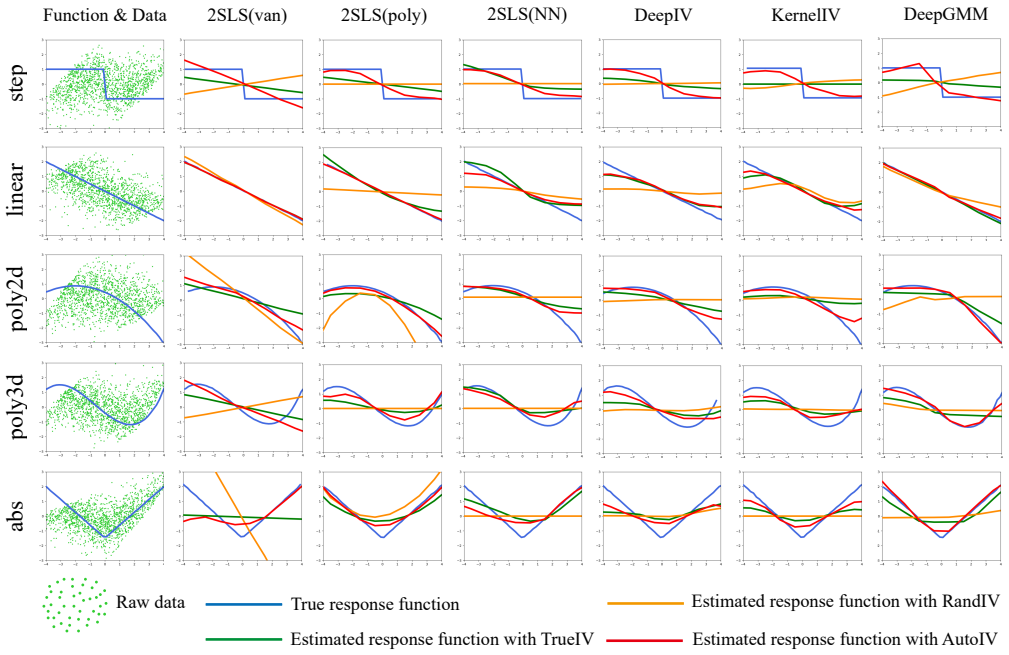

Fig. 2. Response function estimation in low-dimensional scenarios.

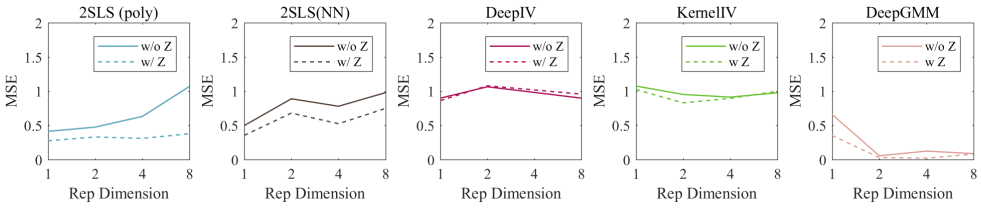

Fig. 3. Performance of AutoIV by varying representation dimensions.

7. **DeepGMM**[3]: a variational method based on optimally-weighted GMM . We use its implementation in CausalML [3].

We compare our algorithm **AutoIV** with the following baseline methods: (1) **TrueIV:** use true valid IVs as a prior; (2) **RandIV:** use random variables (sampled from the same distribution of the true valid IVs) as IVs; (3) **UAS:**[11] use equally weight to synthesize IVs from the IV candidates; (4) **WAS:**[6] synthesize IVs by weighting the IV candidates based on their correlation to the treatment. We use the above methods to generate IVs (IV representations) and feed them to the downstream IV-based counterfactual prediction methods to testify the validity of the generated IVs (IV representations). To evaluate the performance of these IV synthesis methods under different IV candidates validity scenarios, we set: (1) **w/** Z**:** parts of the valid IVs are given in the IV candidates, and (2) **w/o** Z**:** no valid IVs are given in the IV candidates. The latter setting is more practical in real-world applications and would make the task of synthesizing valid IVs (IV representations) more challenging as well as the IV-based counterfactual prediction.

[3]https://github.com/CausalML/DeepGMM

Table 1. Results (MSE±Std) in low-dimensional scenarios over 20 runs.

Methods	IV	step	abs	linear	poly2d	poly3
DirectNN	-	2.86 ± 0.08	2.38 ± 0.05	0.71 ± 0.06	2.34 ± 0.17	1.71 ± 0.07
2SLS (van)	RandIV	2.72 ± 0.82	2.61 ± 0.40	0.33 ± 0.31	1.75 ± 0.80	1.69 ± 0.77
	TrueIV	0.77 ± 0.07	2.05 ± 0.10	0.09 ± 0.06	0.40 ± 0.04	0.25 ± 0.06
	UAS (w/o Z)	3.41 ± 0.17	3.59 ± 0.17	0.95 ± 0.10	3.02 ± 0.23	2.34 ± 0.07
	UAS (w/ Z)	2.15 ± 0.14	2.76 ± 0.12	0.28 ± 0.04	1.76 ± 0.14	1.25 ± 0.08
	WAS (w/o Z)	3.39 ± 0.16	3.58 ± 0.17	0.94 ± 0.10	3.00 ± 0.23	2.32 ± 0.07
	WAS (w/ Z)	2.10 ± 0.18	2.76 ± 0.08	0.28 ± 0.09	1.72 ± 0.18	1.25 ± 0.10
	AutoIV (w/o Z)	1.20 ± 1.96	**1.11 ± 0.09**	**0.00 ± 0.00**	0.22 ± 0.13	0.21 ± 0.06
	AutoIV (w/ Z)	**0.38 ± 0.03**	1.41 ± 0.87	**0.00 ± 0.00**	**0.21 ± 0.08**	**0.19 ± 0.04**
2SLS (poly)	RandIV	2.18 ± 0.70	2.21 ± 0.38	0.87 ± 0.15	1.85 ± 0.71	1.28 ± 0.28
	TrueIV	0.86 ± 0.16	1.96 ± 0.13	0.11 ± 0.07	0.43 ± 0.07	0.28 ± 0.06
	UAS (w/o Z)	3.37 ± 0.14	3.48 ± 0.28	0.98 ± 0.06	3.00 ± 0.18	2.32 ± 0.13
	UAS (w/ Z)	2.11 ± 0.17	2.51 ± 0.31	0.28 ± 0.04	1.63 ± 0.12	1.23 ± 0.09
	WAS (w/o Z)	3.34 ± 0.15	3.55 ± 0.28	0.98 ± 0.06	2.92 ± 0.22	2.30 ± 0.11
	WAS (w/ Z)	2.06 ± 0.19	2.50 ± 0.28	0.28 ± 0.10	1.59 ± 0.17	1.26 ± 0.09
	AutoIV (w/o Z)	**0.39 ± 0.02**	0.41 ± 0.11	**0.00 ± 0.00**	**0.17 ± 0.08**	0.19 ± 0.05
	AutoIV (w/ Z)	**0.39 ± 0.02**	**0.28 ± 0.04**	**0.00 ± 0.00**	0.28 ± 0.21	**0.18 ± 0.04**
2SLS (NN)	RandIV	1.26 ± 0.04	2.09 ± 0.26	0.97 ± 0.05	0.99 ± 0.07	1.02 ± 0.02
	TrueIV	1.04 ± 0.12	1.99 ± 0.20	**0.14 ± 0.02**	0.58 ± 0.06	0.32 ± 0.08
	UAS (w/o Z)	2.46 ± 0.09	3.33 ± 0.32	0.97 ± 0.05	2.19 ± 0.15	2.08 ± 0.06
	UAS (w/ Z)	1.26 ± 0.04	2.12 ± 0.34	0.97 ± 0.05	0.99 ± 0.07	1.02 ± 0.03
	WAS (w/o Z)	2.45 ± 0.06	3.42 ± 0.31	0.97 ± 0.05	2.20 ± 0.12	2.13 ± 0.10
	WAS (w/ Z)	1.82 ± 0.17	2.68 ± 0.20	0.36 ± 0.09	1.61 ± 0.15	1.17 ± 0.10
	AutoIV (w/o Z)	0.47 ± 0.17	0.50 ± 0.18	0.30 ± 0.23	0.50 ± 0.16	0.33 ± 0.16
	AutoIV (w/ Z)	**0.37 ± 0.09**	**0.35 ± 0.06**	0.25 ± 0.09	**0.45 ± 0.30**	**0.26 ± 0.14**
DeepIV	RandIV	1.50 ± 0.09	1.76 ± 0.33	0.90 ± 0.05	1.41 ± 0.11	1.15 ± 0.11
	TrueIV	1.34 ± 0.09	1.69 ± 0.26	0.72 ± 0.05	1.33 ± 0.12	1.01 ± 0.09
	UAS (w/o Z)	1.64 ± 0.11	1.95 ± 0.30	0.93 ± 0.06	1.53 ± 0.21	1.28 ± 0.13
	UAS (w/ Z)	1.59 ± 0.08	1.82 ± 0.24	0.71 ± 0.06	1.37 ± 0.13	1.13 ± 0.07
	WAS (w/o Z)	1.77 ± 0.16	1.81 ± 0.29	0.94 ± 0.07	1.49 ± 0.15	1.34 ± 0.12
	WAS (w/ Z)	1.59 ± 0.12	1.76 ± 0.30	0.70 ± 0.07	1.38 ± 0.12	1.08 ± 0.09
	AutoIV (w/o Z)	**0.66 ± 0.16**	0.90 ± 0.13	0.70 ± 0.18	0.80 ± 0.14	0.86 ± 0.12
	AutoIV (w/ Z)	0.72 ± 0.17	**0.86 ± 0.08**	**0.63 ± 0.11**	**0.71 ± 0.20**	**0.67 ± 0.13**
KernelIV	RandIV	1.55 ± 0.17	4.79 ± 0.13	0.94 ± 0.11	1.04 ± 0.02	1.10 ± 0.15
	TrueIV	1.24 ± 0.11	3.67 ± 0.68	**0.67 ± 0.06**	1.01 ± 0.02	0.99 ± 0.03
	UAS (w/o Z)	3.15 ± 0.28	5.42 ± 0.13	0.92 ± 0.11	2.41 ± 0.14	1.86 ± 0.09
	UAS (w/ Z)	2.37 ± 0.32	5.04 ± 0.21	**0.67 ± 0.06**	1.77 ± 0.07	1.22 ± 0.38
	WAS (w/o Z)	3.22 ± 0.44	5.39 ± 0.16	0.94 ± 0.11	2.48 ± 0.26	1.88 ± 0.09
	WAS (w/ Z)	2.40 ± 0.24	4.81 ± 0.40	**0.67 ± 0.06**	1.77 ± 0.09	1.04 ± 0.02
	AutoIV (w/o Z)	0.93 ± 0.05	1.07 ± 0.05	0.92 ± 0.08	1.03 ± 0.09	0.90 ± 0.36
	AutoIV (w/ Z)	**0.80 ± 0.17**	**0.90 ± 0.04**	0.78 ± 0.09	**0.78 ± 0.27**	**0.89 ± 0.10**
DeepGMM	RandIV	2.03 ± 0.62	2.53 ± 0.31	0.86 ± 0.21	2.16 ± 0.48	1.35 ± 0.43
	TrueIV	1.03 ± 0.10	1.69 ± 0.28	**0.12 ± 0.05**	0.48 ± 0.08	0.32 ± 0.14
	UAS (w/o Z)	3.65 ± 0.13	2.23 ± 0.76	1.00 ± 0.07	3.53 ± 0.65	2.32 ± 0.32
	UAS (w/ Z)	2.23 ± 0.07	2.33 ± 0.60	0.47 ± 0.04	1.75 ± 0.19	1.12 ± 0.14
	WAS (w/o Z)	3.74 ± 0.14	2.21 ± 0.81	1.01 ± 0.07	3.47 ± 0.50	2.31 ± 0.31
	WAS (w/ Z)	2.22 ± 0.18	2.33 ± 0.59	0.43 ± 0.10	1.75 ± 0.18	1.13 ± 0.15
	AutoIV (w/o Z)	0.71 ± 0.36	0.66 ± 0.58	0.44 ± 0.29	0.38 ± 0.48	0.37 ± 0.36
	AutoIV (w/ Z)	**0.69 ± 0.43**	**0.35 ± 0.42**	0.28 ± 0.30	**0.22 ± 0.17**	**0.24 ± 0.17**

Table 2. Ablation experiments of AutoIV.

Methods	$\mathcal{L}_{ZX_m} + \mathcal{L}_{ZY_m}$	$\mathcal{L}_{CX_m} + \mathcal{L}_{CY_m}$	$\mathcal{L}_{CZ_m}$	$\mathcal{L}_{X_r} + \mathcal{L}_{Y_r}$	Results
DeepIV		✓	✓	✓	0.95 ± 0.05
	✓		✓	✓	0.92 ± 0.06
	✓	✓		✓	0.96 ± 0.06
	✓	✓	✓		0.98 ± 0.06
	✓	✓	✓	✓	**0.86 ± 0.08**
KernelIV		✓	✓	✓	0.91 ± 0.09
	✓		✓	✓	0.98 ± 0.11
	✓	✓		✓	1.11 ± 0.15
	✓	✓	✓		> 10
	✓	✓	✓	✓	**0.90 ± 0.04**
DeepGMM		✓	✓	✓	0.49 ± 0.25
	✓		✓	✓	0.60 ± 0.62
	✓	✓		✓	0.68 ± 0.31
	✓	✓	✓		0.73 ± 0.57
	✓	✓	✓	✓	**0.35 ± 0.42**

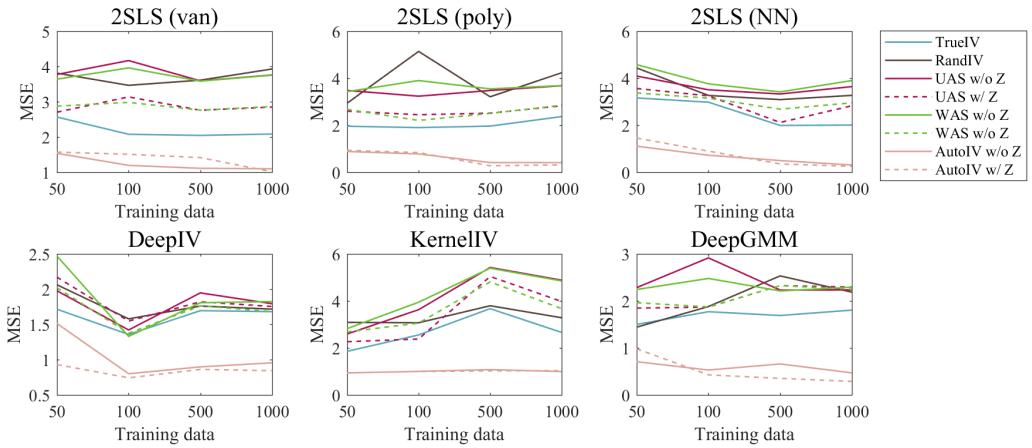

Fig. 4. Performance of AutoIV by varying the training data size.

5.1 Low-dimensional Scenarios

Similar to [3], we first implement experiments in low-dimensional scenarios (i.e., all the variables are in low-dimensional), and the data generating process is:

$$\begin{aligned} &Y = g(X) + e + \sigma,\ X = Z_1 + e + \gamma,\ Z \sim \mathrm{Unif}([-3,3]^2) \\ &V = [Z;\gamma;\sigma], \qquad e \sim \mathcal{N}(0,1), \qquad \gamma,\sigma \sim \mathcal{N}(0,0.1), \end{aligned} \tag{18}$$

where Z are the true valid IVs used as prior in the TrueIV baseline, while RandIV replaces it by randomly sampling from the same distribution of Z. σ and γ are noise. Variables V are observed and used as the IV candidates which is composed by concatenating Z, γ, and σ. e is an unobserved error term that is correlated to both the treatment X and the outcome Y, g is the true response function that chosen from the following settings (some are different from [3] to increase the difficulty of

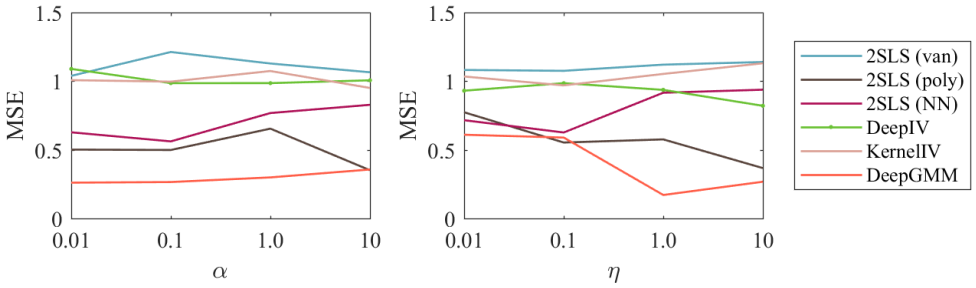

Fig. 5. Results of sensitivity analysis of the hyperparameters α (left) and η (right) in the AutoIV algorithm.

Table 3. Results of high-dimensional experiments on MNIST data with representation dimension as 5, 10, and 15.

Methods	Scenarios	RandIV	TrueIV	AutoIV-5	AutoIV-10	AutoIV-15
2SLS(van)	MNIST_Z	1.688 ± 0.229	**0.986 ± 0.030**	0.994 ± 0.045	1.046 ± 0.055	0.995 ± 0.042
	MNIST_C	1.657 ± 0.170	1.022 ± 0.046	**0.999 ± 0.040**	1.031 ± 0.053	1.018 ± 0.047
	MNIST_{ZC}	2.053 ± 0.307	1.780 ± 0.283	1.006 ± 0.041	1.019 ± 0.040	**0.999 ± 0.033**
2SLS(poly)	MNIST_Z	1.792 ± 1.411	0.977 ± 0.032	**0.444 ± 0.142**	0.604 ± 0.304	0.530 ± 0.207
	MNIST_C	1.491 ± 1.307	0.982 ± 0.041	**0.426 ± 0.053**	0.533 ± 0.207	0.676 ± 0.306
	MNIST_{ZC}	1.327 ± 0.570	1.001 ± 0.027	**0.703 ± 0.219**	0.928 ± 0.065	0.976 ± 0.036
2SLS(NN)	MNIST_Z	1.382 ± 0.110	1.045 ± 0.068	0.663 ± 0.219	0.369 ± 0.066	**0.336 ± 0.043**
	MNIST_C	1.352 ± 0.073	1.074 ± 0.077	0.785 ± 0.196	0.374 ± 0.072	**0.323 ± 0.046**
	MNIST_{ZC}	1.501 ± 0.068	1.427 ± 0.076	0.967 ± 0.081	0.881 ± 0.142	**0.829 ± 0.224**
DeepIV	MNIST_Z	1.102 ± 0.0912	1.030 ± 0.054	**0.875 ± 0.135**	0.891 ± 0.053	0.985 ± 0.117
	MNIST_C	1.221 ± 0.107	1.590 ± 0.402	**0.956 ± 0.118**	1.111 ± 0.144	1.191 ± 0.088
	MNIST_{ZC}	1.163 ± 0.240	1.269 ± 0.336	**1.047 ± 0.033**	1.191 ± 0.119	1.088 ± 0.106
KernelIV	MNIST_Z	0.978 ± 0.034	0.984 ± 0.038	0.968 ± 0.037	0.967 ± 0.034	**0.941 ± 0.044**
	MNIST_C	0.979 ± 0.038	0.979 ± 0.038	**0.960 ± 0.033**	0.972 ± 0.037	0.977 ± 0.034
	MNIST_{ZC}	0.984 ± 0.034	0.984 ± 0.034	**0.944 ± 0.052**	0.966 ± 0.036	0.966 ± 0.036
DeepGMM	MNIST_Z	1.040 ± 0.213	0.586 ± 0.225	0.229 ± 0.333	**0.064 ± 0.091**	0.124 ± 0.227
	MNIST_X	1.108 ± 0.255	0.923 ± 0.086	**0.122 ± 0.182**	0.204 ± 0.309	0.495 ± 0.394
	MNIST_{ZC}	1.051 ± 0.242	0.471 ± 0.129	0.026 ± 0.019	**0.012 ± 0.009**	0.014 ± 0.014

counterfactual prediction):

$$
\begin{aligned}
&\textbf{step}: g(X) = \begin{cases} -1 & X \geq 0 \\ 0 & X < 0 \end{cases} \\
&\textbf{Linear}: g(X) = -X \\
&\textbf{poly2d}: g(X) = -0.1 * X^2 - 0.4 * X \\
&\textbf{poly3d}: g(X) = 0.05 * X^3 + 0.1 * X^2 - 0.8 * X \\
&\textbf{abs}: g(X) = |X|
\end{aligned}
\tag{19}
$$

We sample 500 samples for training, validation, and test, respectively. The values of Z, X, and Y are standardized to avoid the numerical problem. The representation dimensions of Z and C are set to the same, which is a hyper-parameter (the robustness of it is discussed in the later experiments). We plot the true and the estimated response function (i.e., g and $\hat{g}$) in Figure 2. If the IVs fed in

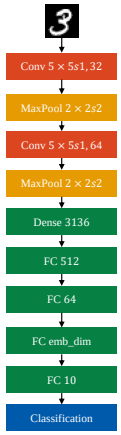

Fig. 6. Convolutional networks for MNIST data. The data are sampled on the penultimate fully-connected (FC) layer.

each method are more valid, the estimated response function would be more closer to the true response function (blue line). We find that (1) RandIV (orange line) fails badly in each case, while TrueIV achieves significantly better performance than RandIV, which indicates that IV information is necessary for removing confounding effect; (2) AutoIV (red line) achieves comparable or even better performance than TrueIV. It is may because AutoIV employs mutual information constraints as well as the representation calibration to further improve the IV representations validity, i.e., enhancing the relevance of the generated IV representations to the treatment and the exclusion to the outcome.

To further improve the difficulty of the task, we then provide a more challenging data generating process by introducing confounders C:

$$\begin{aligned} &Y = g(X) + C_{1...6} + e + \sigma, \quad X = Z_{1...2} + C_{1...6} + e + \gamma \\ &Z \sim \text{Unif}\left([-0.5, 0.5]^2\right), \quad C \sim \text{Unif}\left([-0.5, 0.5]^6\right) \\ &e \sim \mathcal{N}(0,1), \quad \gamma, \sigma \sim \mathcal{N}(0, 0.1), \quad V = [Z; C_{1:4}; \gamma; \sigma] \end{aligned} \tag{20}$$

where $C_{1...6}$ denotes $C_1 + C_2 + ... + C_6$, $Z_{1...2}$ denotes $Z_1 + Z_2$. $C_{1:4}$ is used as a part of IV candidates for IV representation learning, while $C_{5:6}$ is directly employed for the downstream counterfactual prediction methods. We report Mean Square Error (MSE) and standard error (Std) of the predicted counterfactual outcome over 20 runs in Table 1. Similarly, we first find that RandIV performs poorly than TrueIV, indicating that valid IVs are important for removing confounding effect and accurate counterfactual prediction. Besides, the UAS, WAS, and AutoIV methods under w/ Z setting achieve significantly better performance than w/o Z setting, which is probably because the validity of the IV candidates allows IV synthesis methods to generate more valid IV representations. It is worth noting that most of the results under w/o Z setting with AutoIV method show better counterfactual prediction performance even compared with other methods under w/ Z setting. It suggests that AutoIV generates valid IV representations even there is no IV candidate is valid, and we attribute the success to the powerful ability of AutoIV in information control that makes the learned IV representations effectively satisfy the relevance and the exclusion conditions of the valid IVs for accurate counterfactual prediction.

Since representation dimension is a hyperparameter of the AutoIV algorithm, we design experiments by changing the representation dimension as 1, 2, 4, 8 (the true response function g is

Table 4. Results of high-dimensional scenarios on MNIST data with different data composition.

d_Z: 5, d_F: 10, d_A: 4, d_U: 1				
Methods	Scenarios	RandIV	TrueIV	AutoIV
DirectNN	MNIST_Z	1.0314 ± 0.0584	-	-
	MNIST_C	1.3136 ± 0.1213	-	-
	MNIST_{ZC}	1.5746 ± 0.1054	-	-
2SLS(van)	MNIST_Z	1.8015 ± 0.4056	**0.9827 ± 0.0403**	0.9897 ± 0.0469
	MNIST_C	1.6923 ± 0.3306	**0.9880 ± 0.0417**	1.0100 ± 0.0820
	MNIST_{ZC}	1.8393 ± 0.3410	1.4693 ± 0.1897	**1.0064 ± 0.0296**
2SLS(poly)	MNIST_Z	0.9996 ± 0.0573	0.9783 ± 0.0378	**0.4085 ± 0.1414**
	MNIST_C	0.9757 ± 0.0446	0.9666 ± 0.0397	**0.4965 ± 0.1397**
	MNIST_{ZC}	0.9815 ± 0.0325	0.9816 ± 0.0324	**0.8666 ± 0.1465**
2SLS(NN)	MNIST_Z	1.1456 ± 0.1052	0.7765 ± 0.0358	**0.2601 ± 0.0273**
	MNIST_C	1.3150 ± 0.1590	0.8302 ± 0.0693	**0.3698 ± 0.0473**
	MNIST_{ZC}	1.2535 ± 0.0574	1.1636 ± 0.0886	**0.7349 ± 0.2369**
DeepIV	MNIST_Z	1.0356 ± 0.1509	1.2036 ± 0.2065	**0.8174 ± 0.1286**
	MNIST_C	1.1665 ± 0.2003	**1.0873 ± 0.1486**	1.1090 ± 0.1750
	MNIST_{ZC}	1.5355 ± 0.2431	1.2730 ± 0.2021	**1.0168 ± 0.0727**
KernelIV	MNIST_Z	0.9791 ± 0.0392	0.9827 ± 0.0374	**0.9583 ± 0.0477**
	MNIST_C	0.9671 ± 0.0370	0.9671 ± 0.0370	**0.9604 ± 0.0408**
	MNIST_{ZC}	0.9826 ± 0.0339	0.9826 ± 0.0339	**0.9673 ± 0.0351**
DeepGMM	MNIST_Z	0.9679 ± 0.1546	0.5502 ± 0.2183	**0.3052 ± 0.3722**
	MNIST_C	1.1639 ± 0.2048	0.9097 ± 0.0944	**0.1827 ± 0.2632**
	MNIST_{ZC}	1.0206 ± 0.1458	0.4286 ± 0.0566	**0.0074 ± 0.0058**
d_Z: 10, d_F: 5, d_A: 4, d_U: 1				
Methods	Scenarios	RandIV	TrueIV	AutoIV
DirectNN	MNIST_Z	1.3030 ± 0.0706	-	-
	MNIST_C	1.2205 ± 0.0799	-	-
	MNIST_{ZC}	1.8299 ± 0.1313	-	-
2SLS(van)	MNIST_Z	2.0021 ± 0.1951	**0.9894 ± 0.0282**	1.0096 ± 0.0515
	MNIST_C	1.7111 ± 0.1561	**0.9826 ± 0.0474**	1.0242 ± 0.0331
	MNIST_{ZC}	1.9091 ± 0.2820	0.9990 ± 0.0394	**0.9981 ± 0.0543**
2SLS(poly)	MNIST_Z	2.5048 ± 2.3093	0.9878 ± 0.0298	**0.6709 ± 0.2765**
	MNIST_C	2.0190 ± 1.5183	0.9761 ± 0.0440	**0.3794 ± 0.0876**
	MNIST_{ZC}	1.6373 ± 0.6966	0.9692 ± 0.0310	**0.8461 ± 0.1283**
2SLS(NN)	MNIST_Z	1.3283 ± 0.1149	0.8832 ± 0.0789	**0.4526 ± 0.1766**
	MNIST_C	1.1980 ± 0.0629	0.9304 ± 0.0505	**0.3484 ± 0.0229**
	MNIST_{ZC}	1.3864 ± 0.0744	1.0940 ± 0.0645	**0.6992 ± 0.1224**
DeepIV	MNIST_Z	1.1037 ± 0.1253	1.1432 ± 0.1797	**0.9850 ± 0.1336**
	MNIST_C	1.2980 ± 0.0987	1.1619 ± 0.2628	**1.0055 ± 0.1111**
	MNIST_{ZC}	1.1620 ± 0.2386	1.5090 ± 0.4130	**0.9963 ± 0.0733**
KernelIV	MNIST_Z	0.9815 ± 0.0219	0.9839 ± 0.0276	**0.9404 ± 0.0547**
	MNIST_C	0.9771 ± 0.0427	0.9771 ± 0.0427	**0.9613 ± 0.0379**
	MNIST_{ZC}	0.9769 ± 0.0326	0.9769 ± 0.0326	**0.9523 ± 0.0335**
DeepGMM	MNIST_Z	1.0743 ± 0.1008	0.7820 ± 0.2455	**0.0367 ± 0.0535**
	MNIST_C	1.0260 ± 0.1733	0.9021 ± 0.1084	**0.2259 ± 0.2791**
	MNIST_{ZC}	1.1627 ± 0.2468	0.4830 ± 0.1252	**0.0107 ± 0.0082**
d_Z: 10, d_F: 10, d_A: 4, d_U: 1				
Methods	Scenarios	RandIV	TrueIV	AutoIV
DirectNN	MNIST_Z	1.3170 ± 0.0568	-	-
	MNIST_C	1.3577 ± 0.0742	-	-
	MNIST_{ZC}	1.6820 ± 0.1113	-	-
2SLS(van)	MNIST_Z	1.6878 ± 0.2286	**0.9861 ± 0.0297**	1.0464 ± 0.0552
	MNIST_C	1.6570 ± 0.1699	**1.0223 ± 0.0460**	1.0312 ± 0.0527
	MNIST_{ZC}	2.0527 ± 0.3070	1.7801 ± 0.2825	**1.0192 ± 0.0399**
2SLS(poly)	MNIST_Z	1.7924 ± 1.4106	0.9771 ± 0.0315	**0.6043 ± 0.3041**
	MNIST_C	1.4912 ± 1.3071	0.9824 ± 0.0414	**0.5331 ± 0.2065**
	MNIST_{ZC}	1.3272 ± 0.5701	1.0008 ± 0.0267	**0.9277 ± 0.0652**
2SLS(NN)	MNIST_Z	1.3819 ± 0.1103	1.0454 ± 0.0675	**0.3687 ± 0.0656**
	MNIST_C	1.3521 ± 0.0727	1.0743 ± 0.0768	**0.3740 ± 0.0722**
	MNIST_{ZC}	1.5010 ± 0.0678	1.4271 ± 0.0759	**0.8810 ± 0.1422**
DeepIV	MNIST_Z	1.1016 ± 0.0912	1.0304 ± 0.0535	**0.8910 ± 0.0529**
	MNIST_C	1.2205 ± 0.1069	1.5897 ± 0.4022	**1.1111 ± 0.1439**
	MNIST_{ZC}	1.2625 ± 0.2401	1.2693 ± 0.3363	**1.1914 ± 0.1190**
KernelIV	MNIST_Z	0.9777 ± 0.0336	0.9838 ± 0.0383	**0.9668 ± 0.0341**
	MNIST_C	0.9793 ± 0.0377	0.9793 ± 0.0377	**0.9724 ± 0.0368**
	MNIST_{ZC}	0.9842 ± 0.0339	0.9842 ± 0.0339	**0.9658 ± 0.0358**
DeepGMM	MNIST_Z	1.0401 ± 0.2125	0.5864 ± 0.2247	**0.0640 ± 0.0906**
	MNIST_C	1.1075 ± 0.2546	0.9226 ± 0.0857	**0.2038 ± 0.3093**
	MNIST_{ZC}	1.0513 ± 0.2420	0.4711 ± 0.1292	**0.0123 ± 0.0094**

set to abs) and the results are shown in Figure 3. We find that 2SLS (poly) and 2SLS (NN) is not robust enough to the changes of representation dimensions, which is may because their models

are relatively simple. We also see that DeepIV and KernelIV in both w/ Z and w/o Z settings are robust to the representation dimensions. While we note that DeepGMM method performs better in larger representation dimension setting, which is may because DeepGMM relies more heavily on parameter size, and higher dimensions bring more parameters in the fully-connected layer of the neural networks for DeepGMM.

AutoIV is a data-driven decomposed representation learning method, hence we implement experiments with different training data size settings (g is set to abs) as shown in Figure 4. It illustrates that AutoIV achieves great performance in different training data size settings. Moreover, larger data size will increase the decomposed representation learning performance and counterfactual prediction accuracy. However, it is not evident that the performance of other baseline methods is related to the training data size.

We then give sensitivity analysis of the hyperparameters, i.e. α and η in our algorithm. We show the performance of each method in the search space of each hyperparameter in Figure 5. It illustrates that in general the performance of our AutoIV algorithm is robust to α and η with different downstream IV-based methods in counterfactual prediction.

To show the effectiveness of each part of the AutoIV algorithm, we conduct ablation studies by removing each component, including representation learning of Z ($\mathcal{L}_{ZX}^{MI} + \mathcal{L}_{ZY}^{MI}$), representation learning of C ($\mathcal{L}_{CX}^{MI} + \mathcal{L}_{CY}^{MI}$), decomposed regularization ($\mathcal{L}_{CZ}^{MI}$), and counterfactual prediction ($\mathcal{L}_X + \mathcal{L}_Y$). We implement the experiments (g is set to abs) on DeepIV, KernelIV, and DeepGMM, and the results are reported in Table 2. It shows that the necessity of each component in our AutoIV algorithm. Moreover, the two-stage procedure is shown important for further representation calibration. It is because mutual information constraints only control the information flow, but do not effectively enable them to be effective IV representations. While the general two-stage calibration process utilizes the gathered information to further synthesize powerful IV representations.

5.2 High-dimensional Scenarios

Following [3], we then implement experiments in high-dimensional scenarios with hand-written digit datasets MNIST [29]. To further testify the representation learning ability of AutoIV, we consider more complicated data composition that observed variables V contain: (1) IVs Z, (2) confounders F (i.e., variables that are related to X and Y), (3) adjustments A (i.e., variables that are only related to Y), (4) and unconcerned variables U (i.e., variables that are independent of both the treatment X and outcome Y). The data generating process is given as:

$$\begin{aligned} &Y = g(X) + \mathbb{E}[F] + \mathbb{E}[A] + e, \quad X = \mathbb{E}[Z] + \mathbb{E}[F] + e \\ &Z \sim \mathcal{N}(0,1)^{dZ}, \quad F \sim \mathcal{N}(0,1)^{dF}, \quad A \sim \mathcal{N}(0,1)^{dA} \\ &U \sim \mathcal{N}(0,1)^{dU}, \quad C = [F, A, U], \quad V = [Z; C], \quad e \sim \mathcal{N}(0,1), \end{aligned} \tag{21}$$

where dZ, dF, dA, dU are the dimensions of Z, F, A, U respectively. Since UAS and WAS are only valid in the linear setting and are not competent to handle high-dimensional non-linear data, hence we compare RandIV, TrueIV, and AutoIV in the experiments of high-dimensional scenarios. Due to the non-linearity and high-dimension of data increase the difficulty of the task, we only consider w/ Z setting in these experiments. The response function g is set to be abs. We then give the following experimental settings:

$$\begin{aligned} &\textbf{MNIST}_{\mathbf{Z}} : Z \xleftarrow{\text{Conv}} \text{MNSIT}_{\text{rand}} \\ &\textbf{MNIST}_{\mathbf{C}} : C \xleftarrow{\text{Conv}} \text{MNSIT}_{\text{rand}} \\ &\textbf{MNIST}_{\mathbf{ZC}} : Z \xleftarrow{\text{Conv}} \text{MNSIT}_{\text{rand}}, C \xleftarrow{\text{Conv}} \text{MNSIT}_{\text{rand}} \end{aligned} \tag{22}$$

where $\text{MNSIT}_{\text{rand}}$ denotes randomly sampling from MNIST datasets. We adopt convolutional architecture (see Figure 6) to handle original MNIST images by following [3, 18], Z and C are sampled on the penultimate fully-connected (FC) layer with given dimensions.

We sample 1000 data points for training, validation, and test, respectively. We set 10, 10, 4, and 1 for d_Z, d_F, d_A, and d_U, respectively, and let representation dimension be 5 (AutoIV-5), 10 (AutoIV-10), 15 (AutoIV-15). The results are reported in Table 3 with MSE and standard error of 20 runs. We find that the results of each method with AutoIV are significantly better than those with RandIV and superior to those with TrueIV. From the settings of AutoIV-5, AutoIV-10, and AutoIV-15, we see that the performance of AutoIV algorithm is robust to the change of representation dimension, showing its effectiveness in IV representation learning.

We then analyze the performance of AutoIV with different dimensions of data composition and report the results in Table 4. It indicates that AutoIV is competent to generate valid IV representations in different data composition settings. All the experimental settings again show AutoIV's powerful representation learning ability in generating valid IV representation for accurate counterfactual prediction, which is even better than directly using the true valid IVs.

Overall, these results highlight the great decomposed representation learning ability of our AutoIV algorithm in automatically generating the representation serving the role of IVs for accurate IV-based counterfactual prediction.

6 CONCLUSIONS

In this paper, we tackle the problem of decomposing and generating valid IV representations from the observed variables (i.e. the IV candidates). We relax the assumptions and conditions used by previous methods in handling this problem. We propose a novel Automatic Instrumental Variable decomposition (AutoIV) algorithm to decompose and learn valid representations of IVs automatically from the observed variables. We learn the IV representations by employing mutual information constraints, making the learned IV representations satisfy the conditions of the valid IVs in an adversarial game. Extensive empirical results in both low-dimensional and high-dimensional scenarios show the effectiveness of the AutoIV algorithm in generating IV representations and using them for IV-based counterfactual prediction with the downstream methods. The proposed AutoIV algorithm is an important addition to the toolkit of causal inference and IV-based counterfactual prediction.

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China (No. 2018AAA0101900), National Natural Science Foundation of China (No. 61625107, No. 62006207), Key R D Projects of the Ministry of Science and Technology (No. 2020YFC0832500), the Fundamental Research Funds for the Central Universities and Zhejiang Province Natural Science Foundation (No. LQ21F020020).

REFERENCES

[1] Chainarong Amornbunchornvej, E. Zheleva, and Tanya Berger-Wolf. 2021. Variable-lag Granger Causality and Transfer Entropy for Time Series Analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15 (2021), 1 – 30.

[2] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

[3] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems, NeurIPS*. 3564–3574.

[4] Jack Bowden, George Davey Smith, and Stephen Burgess. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology* 44, 2 (2015), 512–525.

[5] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology* 40, 4 (2016), 304–314.

[6] Stephen Burgess, Frank Dudbridge, and Simon G Thompson. 2016. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine* 35, 11 (2016), 1880–1906.

[7] Stephen Burgess and Simon G Thompson. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International journal of epidemiology* 42, 4 (2013), 1134–1144.

[8] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Carin Lawrence. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *International conference on machine learning, ICML*.

[9] Hai H Dam, Hussein A Abbass, Chris Lokan, and Xin Yao. 2007. Neural-based learning classifier systems. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 20, 1 (2007), 26–39.

[10] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. Nonparametric instrumental regression. *Econometrica* 79, 5 (2011), 1541–1565.

[11] Neil M Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. 2015. The many weak instruments problem and Mendelian randomization. *Statistics in medicine* 34, 3 (2015), 454–468.

[12] Oana Frunza, Diana Inkpen, and Thomas Tran. 2010. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE transactions on knowledge and data engineering (TKDE)* 23, 6 (2010), 801–814.

[13] Zekai J Gao, Niketan Pansare, and Christopher Jermaine. 2018. Declarative Parameterizations of User-Defined Functions for Large-Scale Machine Learning and Optimization. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 11 (2018), 2079–2092.

[14] Arthur S Goldberger. 1972. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society* (1972), 979–1001.

[15] Chirok Han. 2008. Detecting invalid instruments using L1-GMM. *Economics Letters* 101, 3 (2008), 285–287.

[16] Lars Peter Hansen. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* (1982), 1029–1054.

[17] Jason Hartford, Victor Veitch, Dhanya Sridhar, and Kevin Leyton-Brown. 2020. Valid Causal Inference with (Some) Invalid Instruments. *arXiv preprint arXiv:2006.11386* (2020).

[18] Jason S. Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A Flexible Approach for Counterfactual Prediction. In *International Conference on Machine Learning, ICML*. 1414–1423.

[19] Negar Hassanpour and Russell Greiner. 2020. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations, ICLR*.

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

[21] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. 2016. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association* 111, 513 (2016), 132–144.

[22] Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. 2018. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 8 (2018), 1544–1554.

[23] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable prediction across unknown environments. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. 1617–1626.

[24] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Yashen Wang, Fei Wu, and Shiqiang Yang. 2020. Treatment Effect Estimation via Differentiated Confounder Balancing and Regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14 (2020), 1 – 25.

[25] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *International Conference on Knowledge Discovery and Data Mining, SIGKDD*. 265–274.

[26] Kun Kuang, Peng Cui, Hao Zou, Bo Li, Jianrong Tao, Fei Wu, and Shiqiang Yang. 2020. Data-Driven Variable Decomposition for Treatment Effect Estimation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2020).

[27] Kun Kuang, Hengtao Zhang, Runze Wu, Fei Wu, Yueting Zhuang, and Aijun Zhang. 2021. Balance-Subsampled Stable Prediction Across Unknown Test Data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 3 (2021), 1–21.

[28] Zhaobin Kuang, Frederic Sala, Nimit Sohoni, Sen Wu, Aldo Córdova-Palomera, Jared Dunnmon, James Priest, and Christopher Ré. 2020. Ivy: Instrumental Variable Synthesis for Causal Inference. *arXiv preprint arXiv:2004.05316* (2020).

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[30] Greg Lewis and Vasilis Syrgkanis. 2018. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164* (2018).
[31] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. 2019. Dual iv: A single stage instrumental variable regression. *arXiv preprint arXiv:1910.12358* (2019).
[32] Whitney K Newey and James L Powell. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 5 (2003), 1565–1578.
[33] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12700–12710.
[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
[35] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2019).
[36] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning, ICML*. 3076–3085.
[37] Rahul Singh, Maneesh Sahani, and Arthur Gretton. 2019. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems, NeurIPS*. 4593–4605.
[38] Hao Wang and Dit-Yan Yeung. 2016. Towards Bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 28, 12 (2016), 3395–3408.
[39] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 3091–3100.
[40] Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. 2019. On the use of the lasso for instrumental variables estimation with some invalid instruments. *J. Amer. Statist. Assoc.* 114, 527 (2019), 1339–1350.
[41] Jeffrey M Wooldridge. 2002. Econometric analysis of cross section and panel data MIT press. *Cambridge, MA* 108 (2002).
[42] Jeffrey M Wooldridge. 2016. Should instrumental variables be used as matching variables? *Research in Economics* 70, 2 (2016), 232–237.
[43] Philip G Wright. 1928. *Tariff on animal and vegetable oils*. Macmillan Company, New York.
[44] Anpeng Wu, Kun Kuang, Junkun Yuan, Bo Li, Pan Zhou, Jianrong Tao, Qiang Zhu, Yueting Zhuang, and Fei Wu. 2020. Learning Decomposed Representation for Counterfactual Inference. *arXiv preprint arXiv:2006.07040* (2020).
[45] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9593–9602.
[46] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and A. Zhang. 2021. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15 (2021), 1 – 46.
[47] Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. 2019. On the Estimation of Treatment Effect with Text Covariates.. In *International Joint Conference on Artificial Intelligence, IJCAI*. 4106–4113.
[48] Kui Yu, L. Liu, and Jiuyong Li. 2021. A Unified View of Causal and Non-causal Feature Selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15 (2021), 1 – 46.
[49] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2021. Transporting Causal Mechanisms for Unsupervised Domain Adaptation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 8599–8608.