



Estimating Individualized Causal Effect with Confounded Instruments

Haotian Wang

Institute for Quantum Information & State Key
Laboratory of High Performance Computing,
College of Computer, National University of
Defense Technology
wanghaotian13@nudt.edu.cn

Wenjing Yang*

Institute for Quantum Information & State Key
Laboratory of High Performance Computing,
College of Computer, National University of
Defense Technology
wenjing.yang@nudt.edu.cn

Longqi Yang

Institute for Quantum Information & State Key
Laboratory of High Performance Computing,
College of Computer, National University of
Defense Technology
yanglongqi19@nudt.edu.cn

Anpeng Wu

Institute of Artificial Intelligence,
Zhejiang University
anpwu@zju.edu.cn

Liyang Xu

Institute for Quantum Information & State Key
Laboratory of High Performance Computing,
College of Computer, National University of
Defense Technology
xuliyang08@nudt.edu.cn

Jing Ren

Institute for Quantum Information & State Key
Laboratory of High Performance Computing,
College of Computer, National University of
Defense Technology
renjing@nudt.edu.cn

Fei Wu[†]

Institute of Artificial Intelligence,
Zhejiang University
wufei@zju.edu.cn

Kun Kuang[‡]

Institute of Artificial Intelligence,
Zhejiang University
kunkuang@zju.edu.cn

ABSTRACT

Learning individualized causal effect (ICE) plays a vital role in various fields of big data analysis, ranging from fine-grained policy evaluation to personalized treatment development. However, the presence of unmeasured confounders increases the difficulty of estimating ICE in real-world scenarios. A wide range of methods have been proposed to address the unmeasured confounders with the aid of instrument variable (IV), which sources from the treatment randomization. The performance of these methods relies on the well-predefined IVs that satisfy the unconfounded instruments assumption (i.e., the IVs are independent with the unmeasured confounders given observed covariates), which is untestable and leads to finding a valid IV becomes an art rather than science. In this paper, we focus on estimating the ICE with *confounded instruments* that violate the unconfounded instruments assumption. By considering the conditional independence between the set of confounded instruments and the outcome variable, we propose a novel method, named *CVAE-IV*, to generate a substitute of the *unmeasured confounder* with a conditional variational autoencoder. Our theoretical analysis guarantees that the generated confounder substitute will identify unbiased ICE. Extensive experiments on bias demand prediction and Mendelian randomization analysis verify the effectiveness of our method.

*Wenjing Yang and Haotian Wang contributed equally to this research.

[†]Shanghai Institute for Advanced Study of Zhejiang University; Shanghai AI Laboratory

[‡]Kun Kuang is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '22, August 14–18, 2022, Washington, DC, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9385-0/22/08...\$15.00

<https://doi.org/10.1145/3534678.3539335>

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; *Supervised learning by regression*; • **Mathematics of computing** → *Causal networks*.

KEYWORDS

Individualized Causal Effect, Instrument Variable

ACM Reference Format:

Haotian Wang, Wenjing Yang, Longqi Yang, Anpeng Wu, Liyang Xu, Jing Ren, Fei Wu, and Kun Kuang. 2022. Estimating Individualized Causal Effect with Confounded Instruments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3534678.3539335>

1 INTRODUCTION

The popularity of big data has improved the ability of machine learning methods to inform actions to be taken in the physical world. However, due to the requirement for robust and reliable performance, decision makers often prefer to ask counterfactual questions on the individualized causal effect (ICE) that an action will have before it is taken. The availability of large training datasets has increased research interest in inferring ICE from observational data [14, 24]; examples include the application of uplift modeling to advertising recommendation [14] or demand analysis to airline price [6]. In more detail, Fig. 1 illustrates the case of airline demand analysis [6], in which decision makers aim to determine the ICE of the airline price T (treatment) on the sales tendency Y (outcome) given personalized characteristic X (e.g., specific holidays). However, the estimation of ICE from observational data becomes generally impossible due to the presence of unmeasured confounders E (people's demands, such as big conferences), as such variables simultaneously affect the treatment and the outcome [11].

With the aid of instrument variable (IV), one can isolate the estimation on ICE from the influence of unmeasured confounders [6]. For instance, assuming that one can access the fuel cost Z , which is independent of will create's demand (unmeasured confounder) and affects the sales tendency only through the price, then the change of fuel cost will create movement in airline prices independent of latent ticket demand and further enables

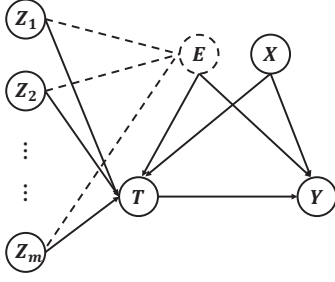


Figure 1: Causal graph of our air-travel demand example, where the treatment T , outcome Y and observed covariates X represent airline price, ticket sales and holiday recordings, respectively. The unmeasured confounders E represent people’s demand for tickets or the fluctuation of the economic situation. The set of IV candidates $\{Z_i\}_{i=1}^m$ represents the fuel cost recorded in different periods. The solid arrow represents a direct causal relationship, while the dashed line represents a (possible) indirect correlation.

causal effect estimation. A large number of methods have subsequently been developed that take the accessibility of valid instrument variables as a basic premise, including two-stage regression [6], generalized moment methods [2], control functions [22] and dual formulation [20].

However, one of the most crucial aspects of applying IV methods in real-world cases is to first find a valid IV [7]. Specifically, a valid IV should satisfy three key assumptions. The *relevance* assumption requires that an IV should be correlated with the treatment variable when the observed covariates X is given, which can be easily verified using observational data. The *exclusion* assumption asserts that the IV should affect the outcome only through the treatment. Although directly verifying the exclusion assumption is difficult, it is still possible to rule out anti-exclusion IVs using robust estimators [9]. Different from the relevance and exclusion assumptions, which hold that all relevant variables are measured, the *Unconfounded Instruments* assumption remains the key challenge for defining a valid IV, which states that the unmeasured confounder and instruments are conditionally independent given observed covariates X . Due to the presence of unmeasured confounders, testing the independence between IVs and confounders becomes problematic. As an instance, we add the economic fluctuations into the category of unmeasured confounders E in Fig. 1, since these economic fluctuations affect both the airline company’s ticket prices and the customers’ willingness to purchase tickets. Consequently, Z_i serves as a valid IV if its corresponding fuel cost is recorded before the period of the economic fluctuation. On the other hand, Z_i is invalid if its fuel cost are recorded during the economic fluctuation period, due to its correlation with the unmeasured confounder E . To filter out such invalid IVs, one can only rely on the prior from domain experts with sophisticated designs.

In this paper, in the presence of unmeasured confounders, we concentrate on identifying causal effects with a bunch of IV candidates, which becomes a paradigm across diverse domain applications (e.g., multiple genetic markers in Mendelian randomization [9] or randomized trial judges in economics [7]). As shown in Fig. 1, we allow the existence of “confounded IV candidates” without the pre-knowledge that which IV candidate is invalid, where confounded IVs only violate the unconfounded assumption ($Z_i \not\perp E | X$) but satisfy both the relevance and exclusion assumptions. Our presented setting is both practical and commonly inevitable when domain knowledge is vague. Herein, we aim to develop a general approach to identify non-linear individualized causal effect under violation of the unconfounded instrument assumption.

Rather than recommending the modeler to wait for inspiration from domain experts, or to simply gamble by selecting one IV candidate, we instead

propose to learn a “valid” substitute of unmeasured confounders with the aid of the entire set of IV candidates. More specifically, by constructing a deep conditional variational autoencoder (CVAE-IV), we aim to generate a substitute E' of unmeasured confounder E that obeys strong ignorability, such that $Y_t \perp T | E', X$. To achieve the ignorability, we design the CVAE-IV under the statistical principle that $Y \perp \{Z_i\}_{i=1}^m | T, E', X$, which states that the outcome and IV candidates are conditionally independent given the treatment, observed covariates and the generated E' . Our theoretical analysis guarantees that once the CVAE-IV is well optimized, the generated substitute E' is sufficient to support unbiased individualized causal effect identification. Unlike previous work [7, 9, 15] which mainly focuses on invalid IVs that violates the exclusion assumption, we challenge the unconfounded instrument assumption with more general identification results on non-linear ICE.

In summary, we highlight our contributions as follows:

- We challenge the estimation of non-linear individualized causal effect with confounded IV candidates in the presence of unmeasured confounders;
- We propose a novel method, named CVAE-IV, to construct ignorable confounder substitute for predicting ICE. Meanwhile, we provide genetic theoretical guarantee for unbiased identification on ICE by our method;
- We conduct extensive experiments on biased demand prediction with both low-dimensional and high-dimensional features, and the realistic Mendelian randomization analysis with hundreds of IV candidates. Experimental results verify that our method successfully isolates the estimation on ICE from the confounder, with superior prediction results.

2 RELATED WORK

2.1 Background on Instrument Variables

To remove the influence of the unmeasured confounder for identifying individualized causal effect (ICE) in observational study [11, 14], the most popular approach involves the accessibility of valid instrument variables (IV) [11]. Under the typical separable assumption [6, 11], the identification of the ICE reduces to solving the Fredholm integral equation [6] with completeness condition [22]. A variety of approaches to achieving this goal have been proposed [1], with the most classical solutions including include the two-stage regression (2SLS) [1] and Wald estimator [1]. Following this regime, methods such as DeepIV [6] and KernelIV [25] have been proposed as non-linear expansions. Meanwhile, generalized method of moments (GMM) [2] constitutes another direction by using of nonlinear basis functions to identify ICE. Furthermore, the dual formulation has also been applied together with variational inference [20] have also been applied to obtain IV solutions. In this paper, we focus on the “weaker” separable setting in which the outcome structural equation is still additive but imposes no constraint on the confounder’s mean.

2.2 Inference with Invalid Instruments

Most existing related research concentrates on the setting with the violation of exclusion assumption [7, 9]. Such methods often migrate the bias through averaging over the bias or eliminating the bias using concepts drawn from robust statistics [9]. Especially, a recent work named Mod-eIV proposed to ensemble multiple IV solutions based on the modal validity [7]. For broader violations of both the exclusion and unconfounded assumptions, existing methods must rely on the prior fact that the majority of IV candidates are valid [15]. Consequently, they treat biased instruments as outliers and apply robust estimator to eliminate them [5]. However, these methods often assume the linear and homogeneous causal effect [13] or prior structural information [15]. Unfortunately, the former requirement violates the ICE estimation, while the latter condition is not

suitable for our setting by requiring the prior knowledge that which IV candidate is invalid. Aside from the separable case, recent semi-parametric inference methods [8, 16] have achieved the unbiased identification of non-linear ICE under the valid majority assumption with some parametric assumptions (e.g., linear correlation between treatment and the confounder and the dimension reduction assumption [16]). In addition, to relax the IV assumptions, researchers have explored the conditional IVs from the pre-learned partial causal graph [3].

2.3 Negative Control and Proxy Variable

Aside from the IV methods, identifying ICE through observational proxy variables is another approach that has attracted increasing attention in recent years [17, 19]. As a representative example, the double negative control (NC) method provides genetic identification results with the aid of a negative exposure variable (NCE) and a negative outcome variable (NCO) [19]. Recently, some deep learning methods such as CEVAE [17] adopt a variational model to model the joint distribution of both observational and latent variables, while their method is lack of theoretical guarantee. Interestingly, our definition on confounded IV that violates the unconfounded assumption, is similar to that of negative exposure variable [19]. However, our method can be meaningfully distinguished from proxy variable methods, due to the fact that genetic identification of causal effect in proxy variable methods requires the accessibility of both NCE and NCO, while our methods only access to one side (invalid IVs). Although proxy variable method also allows for the identification of ICE using solely NCE, its strong assumption, which states that the linear correlation between NCE and unmeasured confounder is exactly equal to that between treatment and unmeasured confounder, is impractical in real-world applications.

3 PROBLEM AND PRELIMINARY

In this paper, we define causal effects using the potential outcome framework [10]. More specifically, we aim to estimate the individualized causal effect (ICE) of treatment variable $T \in \mathcal{R}^n$ on outcome variable $Y \in \mathcal{R}^n$, conditioned on observed covariates $X \in \mathcal{R}^{n \times d}$ that represents personalized features, where n denotes the sample size, d denotes the personalized feature size. An illustrative example is shown in Fig. 1. Moreover, due to the presence of unmeasured confounders E , we have access to a set of m instrument variables (IV) candidates $\{Z_i\}_{i=1}^m$ [7, 9], where each $Z_i \in \mathcal{R}^n$ is valid if and only if it satisfies:

- (1) Relevance: $T \not\perp Z_i \mid X$;
- (2) Exclusion: Z_i affects Y only through T ;
- (3) Unconfounded Instrument: $Z_i \perp E \mid X$.

On the contrary, Z_i is invalid if the unconfounded instrument assumption is violated; we define such an IV candidate as “confounded IV” (invalid IV) in this paper: $Z_i \not\perp E \mid X$, while relevance and exclusion assumptions still hold for Z_i . In a realistic scenario, we assume that a subset of $\{Z_i\}_{i=1}^m$ are invalid IVs without any prior knowledge stating that which IV candidate is invalid [7]. For instance, recalling the example in Fig. 1, we have recorded fuel cost across multiple periods, but it remains vague that which one coincides with the economic fluctuation. *Throughout this paper, we sometimes refer to \bar{Z} instead of $\{Z_i\}_{i=1}^m$ for the sake of brevity.* In order to guarantee identification of ICE via IVs, we follow the widely used separable assumption [10], which states that the effect of the unmeasured confounders (E) on Y is additive. More formally, we assume that:

$$Y = g_1(X, T) + g_2(E). \quad (1)$$

Throughout this paper, we formulate the causal effect under the potential outcome framework [11], which uses the notation $Y(t)$ to denote the counterfactual outcome of Y with the intervention $T = t$. Hence, the ICE that represents our goal can be written as follows:

$$\mathbb{E}[Y(t)|X] - \mathbb{E}[Y(t')|X]. \quad (2)$$

Algorithm 1 Training CVAE-IV for counterfactual inference

Require: Individualized features X , observed outcome Y , actual treatment $T = t$ and IV candidates $\{Z_i\}_{i=1}^m$
Ensure: Individualized causal effect with intervention $T = t'$.

- 1: **Training procedure:**
- 2: Train CVAE-IV model $\{\cdot, \cdot\}$ by optimizing (6).
- 3: Train regression model $= \{\psi_1, \psi_2\}$ by optimizing (7).
- 4: **Inference procedure:**
- 5: Recover substitute of unmeasured confounder E' from the encoder: $E' \sim p_\phi(E' | Y, \{Z_i\}_{i=1}^m, T, X)$.
- 6: Predict $Y(t')$ based on E' using (8).
- 7: **return** ICE as $Y(t') - Y$ for each individual.

Beyond identifying the ICE with a valid IV Z^* , previous studies have directly identify the effect function g_1 with another assumption that $\mathbb{E}[g_2(E) | X] = 0$. However, once $Z^* \not\perp E | X$ is violated, the endogenous noise term $\mathbb{E}[g_2(E) | X, Z_i]$ biases the estimate of underlying effect function g_1 [2, 6]. Therefore, naively applying IV methods with confounded IVs inevitably bias the estimation of ICE.

4 METHODS

As analyzed above, it is necessary to design a new strategy towards invalid IVs that violate the unconfounded assumption. However, due to the presence of endogenous noise term $\mathbb{E}[g_2(E) | X, Z_i]$, it is impossible to separate the underlying effect function g_1 from the unmeasured confounder [6]. Analogously, the dependence between $\{Z_i\}_{i=1}^m$ and T impedes us from recovering the accurate distribution of unmeasured confounder E in the same way as most proxy variable approaches [17]. Hence, we aim to develop a fallback solution by accurately identifying the variation of the underlying effect function $g_1(X, T)$ when T varies, which is exactly the individualized causal effect $\mathbb{E}[Y(t)|X] - \mathbb{E}[Y(t')|X]$.

The complete framework of our method, which is illustrated in Fig. 2, is divided into three components:

- (a) Generation of confounder substitute: We construct a conditional variational model to learn the substitute E' of unmeasured confounder E from $\{Z_i\}_{i=1}^m, Y$;
- (b) Outcome regression: We fit Y with X, T and generated E' using a parameterized regressor;
- (c) Counterfactual inference: we estimate the counterfactual outcome $Y(t')$ based on learned E' and intervention $T = t$.

In the interests of clarity, we present the procedure of model training together with ICE estimation in Algorithm 1.

4.1 CVAE-IV: Conditional Variational Autoencoder for Invalid IVs

In the first stage, we propose to generate an “ignorable” substitute E' of the unmeasured confounder E with the aid of IV candidates $\{Z_i\}_{i=1}^m$; here, we refer to a confounder substitute as “ignorable” only if it satisfies the strong ignorability criterion [11]: $Y(t) \perp T | E', X$. In order to learn this ignorable substitute, we force the generated E' to satisfy the following principle:

$$Y \perp (Z_1, Z_2, \dots, Z_m) | E', T, X. \quad (3)$$

As shown in Fig. 1, following the learning of the substitute E' , which conditionally cuts off the correlation between Y and $\{Z_i\}_{i=1}^m$ given (X, T) , E' naturally captures all the unobserved confounders E . Presenting a contradiction may be more intuitive: if some component of E is not captured by E' , then such component still “activates” some correlation paths [18] between Y and $\{Z_i\}_{i=1}^m$ through T , resulting in the violation of principle (3). Therefore, strong ignorability criterion holds for E' . We leave the strict theoretical proof behind this intuition for the next section.

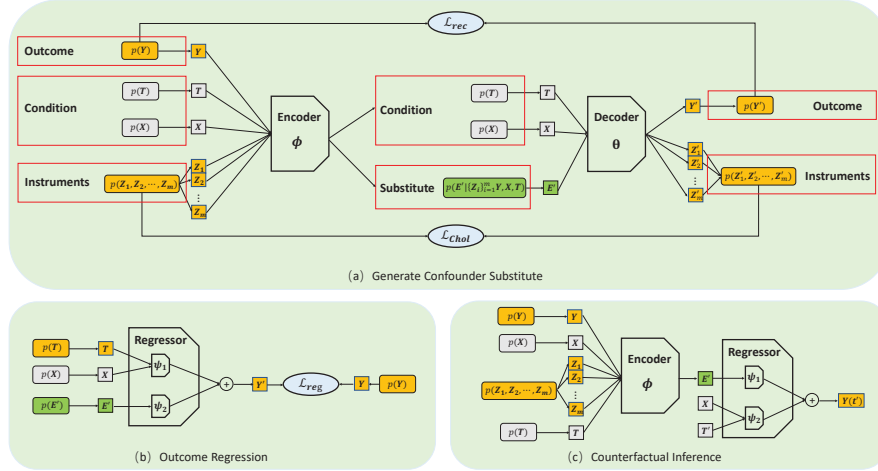


Figure 2: Framework of CVAE-IV. (a). Generating the confounder substitute, (b). Outcome regression, (c). Counterfactual Inference.

Overall, based on the above analysis, it is natural for us to construct a conditional variational autoencoder named “CVAE-IV” to generate the confounder substitute E' . More specifically, we apply the variational inference to model the conditional distribution $p(Y, \bar{Z} = \{Z_i\}_{i=1}^m | T, X)$ as follows [26]:

$$\begin{aligned} & \log p(Y, \bar{Z} | T, X) \\ & \geq \mathbb{E}[\log p_\theta(Y, \bar{Z} | E', T, X)] - D_{KL}(q_\phi(E' | Y, \bar{Z}, X, T) \| p(E' | T, X)), \end{aligned}$$

where \bar{Z} refers to the joint of $\{Z_i\}_{i=1}^m$, D_{KL} refers to the KL-divergence [17] between variational posterior and the underlying one, p_θ is the decoder model and q_ϕ is the encoder model. By forcing the underlying posterior $p(E' | T, X)$ to follow the normal distribution, we obtain the framework of CVAE-IV in Fig. 2 (a).

Importantly, the classical reconstruction loss of VAE or CVAE assumes a diagonal covariance matrix for reconstructed output [26], which implies that such models implicitly assume the pairwise conditional independence of $Y, \{Z_i\}_{i=1}^m$. On the one hand, conditional independence between Y and the joint distribution $\{Z_i\}_{i=1}^m$ given E', T, X exactly coincides with our conditional independence principle (3). On the other hand, however, the internal pairwise conditional independence among IV candidates $\{Z_i\}_{i=1}^m$ is not necessary; in other words, we cannot guarantee that there is no direct/indirect causal relationship between every pair of IV candidates. For example, in the airline demand case, we cannot guarantee that the fuel cost in some period has no effect on that in the following periods. Therefore, based on the conditional principle in (3), we separate the reconstruction of the outcome from that of IV candidates:

$$\log p_\theta(Y, \bar{Z} | E', T, X) = \log p_\theta(Y | E', T, X) + \log p_\theta(\bar{Z} | E', T, X),$$

where the former reconstruction loss is easily obtained:

$$\mathcal{L}_{rec} = \log p_\theta(Y | E', T, X) \approx \frac{1}{\sigma_Y^2} \|Y - Y'(E', T, X)\|_2^2 \quad (4)$$

where $Y'(E', T, X)$ is the reconstructed outcome based on E' and σ_Y is the variance of Y . For the latter reconstruction loss on IV candidates $\log p_\theta(\{Z_i\}_{i=1}^m | E', T, X)$, we should preserve their pairwise dependence by combining the covariance matrix of $\bar{Z} = \{Z_i\}_{i=1}^m$ into the decoder, as follows:

$$\mathcal{L}_{chol} = \log |\Sigma_{\bar{Z}}(E')| + (\bar{Z} - \mu(E'))^T \Sigma_{\bar{Z}}(E')^{-1} (\bar{Z} - \mu(E')), \quad (5)$$

where $\mu(E')$ refers to the mean embedding of $\{Z_i\}_{i=1}^m$, and $\Sigma_{\bar{Z}}(E')$ refers to the covariance matrix of $\{Z_i\}_{i=1}^m$ in the decoder. To facilitate computation,

we follow [4] and turn to estimate the matrix $\Lambda(E') = \Sigma_{\bar{Z}}(E')^{-1}$. Moreover, $\Lambda(E')$ is presented using a Cholesky decomposition trick such that $\Lambda(E') = L(E')L(E')^T$; here, $L(E')$ is a lower triangular matrix and we only estimate $L(E')$ instead of $\Sigma_{\bar{Z}}(E')$ from E' in the decoder. With Cholesky decomposition, the original reconstruction loss \mathcal{L}_{chol} for IV candidates in (5) becomes:

$$\mathcal{L}_{chol} = -2 \sum_{i=1}^m (\log L(E')_{ii}) + (\bar{Z} - \mu(E'))^T L(E')L(E')^T (\bar{Z} - \mu(E'))$$

where $L(E')_{ii}$ refers to the i -th element in the diagonal of $L(E')$. In summary, the CVAE-IV model comprises three parts:

- (1) \mathcal{L}_{chol} represents the reconstruction of IV candidates $\{Z_i\}_{i=1}^m$ by using of $\mu(E')$ and $L(E')$ generated from the latent E' ;
- (2) \mathcal{L}_{rec} represents the reconstruction of outcome Y by using of $\mu(E')$ and $L(E')$ generated from the latent E' ;
- (3) \mathcal{L}_{KL} represents the KL divergence between the variational posterior $q_\phi(E' | Y, \bar{Z}, T)$ and the normal distribution.

Thus, the overall objective for optimizing CVAE-IV is:

$$\mathcal{L}_{gen} = \mathcal{L}_{chol} + \mathcal{L}_{rec} + \lambda * \mathcal{L}_{KL}, \quad (6)$$

where the λ controls the variance of the reconstructed output.

4.2 Outcome Regression and Counterfactual Inference

Once the CVAE-IV model is optimized, we fit the conditional distribution of the observational outcome Y $p_\phi(Y | X, T, E')$ as follows:

$$\mathcal{L}_{reg} = \|Y - (g_{\psi_1}(X, T) + g_{\psi_2}(E'))\|_2^2, \quad (7)$$

where the regression functions g_{ψ_1} and g_{ψ_2} are parametrized by deep networks with ψ_1 and ψ_2 . It is noteworthy that we use the separation assumption as a known prior, which implies that we explicitly assume the effect of unmeasured confounder E on Y to be additive. Therefore, we plug this inductive bias into our formulation of regression model. More importantly, this formulation plays a vital role in guaranteeing unbiased estimation of ICE, as shown in the next section.

Ultimately, our goal is the accurate prediction of ICE with intervention $T = t'$. As shown in Fig. 2 (c), we first recover the substitute of unmeasured confounder E' from the encoder Φ of trained CVAE-IV, with observed $X, T, \{Z_i\}_{i=1}^m, Y$ (here Y is the observed outcome that is consistent with the

potential outcome $Y(t)$). We then predict the counterfactual outcome $Y(t')$ with the trained regression model $= \{\psi_1, \psi_2\}$:

$$Y(t') = g_{\psi_1}(X, T = t') + g_{\psi_2}(E'), \quad (8)$$

where we finally calculate the ICE in (2) by comparing $Y(t')$ with the observed outcome $Y(t)$.

4.3 Discussion on Substitute Generation

Two key problems might now arise regarding the generation of the confounder substitute in our method:

- (1) Does the learned substitute E' coincide exactly with the underlying E ? No—rather than requiring that $E' = E$ exactly holds, we only require that E' captures E , which further entails that the strong ignorability on E' holds.
- (2) Comparison between our method and previous IV method. We sacrifice strength of the identification results to compensate for the presence of confounded IVs. IV methods exactly identify the causal effect function g_1 , while our method only identifies the variation of g_1 with intervention on T .

5 THEORETICAL ANALYSIS

Beyond intuitive analysis, we provide theoretical proof to justify our key conditional independence principle in (3) in this section. Due to space constraints, detailed proof for all the theorems is provided in the appendix.

5.1 Ignorability of Confounder Substitute

Throughout this subsection, we omit the observed covariate X in the encoder input to generate E' for convenience without loss of generality. We begin our derivation by modifying the “Kallenburg Construction” in [27]:

Definition 5.1. (Conditional Kallenburg Construction) Assume we have a basic probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Consider a random variable E' with domain \mathcal{E}' , then the distribution of $\{(Z_1, Z_2, \dots, Z_m), Y, T\}$ admits a Kallenburg construction if there exists a (deterministic) measurable function $f : (\mathcal{E}' \times \mathcal{T}) \times [0, 1] \rightarrow \mathcal{Z}^m$, random variables $U \in [0, 1]$, measurable function $g : (\mathcal{E}' \times \mathcal{T}) \times [0, 1] \rightarrow \mathcal{Y}$ and random variables $V \in [0, 1]$ such that

$$(Z_1, Z_2, \dots, Z_m) = f((E', T), U) \quad Y = g((E', T), V), \quad (9)$$

where \mathcal{Z}, \mathcal{T} and \mathcal{E} are domains of instrument \bar{Z} , T and E' , respectively. Meanwhile, each of the $\{U, V\}$ marginally follows the uniform distribution $U(0, 1)$ and are jointly independent.

We go on to define the conditional variational autoencoder by separating the parameter of decoder $\Theta = \{\theta_y, \theta_z\}$:

Definition 5.2. (Conditional Variational Autoencoder) Given the observed multiple instruments $\{Z\}_{i=1}^m$, treatment variable T , latent variables E' and a set of parameters $\{\theta_z, \theta_y\}$, a conditional variational autoencoder (CVAE) model $(\{Z\}_{i=1}^m, \{\theta_z, \theta_y\}, E', T)$ defines the joint probability of observations $p_{\theta_z, \theta_y}(\{(Z_1, Z_2, \dots, Z_m), Y\} | T)$:

$$\int p(E' | T) p_{\theta_z}((Z_1, Z_2, \dots, Z_m) | E', T) p_{\theta_y}(Y | E', T) dE', \quad (10)$$

where the posterior distributions $p(E' | T)$ is modeled as a normal Gaussian distribution.

Since the conditional distributions $p_{\theta_z}((Z_1, Z_2, \dots, Z_m) | E', T)$ and $p_{\theta_y}(Y | E', T)$ are modeled via the independently updated θ_z and θ_y , we connect the conditional Kallenburg construction to CVAE model in the following theorem via the weak regularity condition:

Assumption 5.3. (Weak regularity condition.) Consider a random variable Z from basic space $(\Omega, \mathcal{F}, \mathbf{P})$ to its domain \mathcal{Z} , Z satisfies the weak regularity condition if and only if \mathcal{Z} is a Borel space.

Theorem 5.4. Every CVAE Model $(\{Z\}_{i=1}^m, \{\theta_z, \theta_y\}, E', T)$ admits a conditional Kallenburg construction under weak regularity condition.

We then conclude the following property of the CVAE model:

Theorem 5.5. Any hidden confounder E must be captured by the learned CVAE model through the generated substitute E' .

Now it is necessary to introduce a consistency condition [27] to constrain the relationship between the input of CVAE $\{\bar{Z} = \{Z\}_{i=1}^m, Y, T\}$ and the learned substitute E' as follows:

Assumption 5.6. (Consistency) The estimation of the CVAE model is consistent if and only if, for some function f_ϕ ,

$$p(E' | \bar{Z}, Y, T) = \delta(f_\phi(\bar{Z}, Y, T)). \quad (11)$$

An intuitive interpretation of the consistency assumption is that the confounder substitute can be deterministically captured by $\{\bar{Z}, Y, T\}$, which is easily satisfied when we sample E' deterministically from the mean embedding $\mu(E')$. The following theorem therefore holds with the consistency condition:

Theorem 5.7. Any post-treatment variable V of T must not be measured with the generated confounder substitute E' .

Based on our Theorem 5.5 and Theorem 5.7, we can immediately conclude that the learned substitute confounder E' obeys the well-known “back-door” criterion [18] that E' captures every confounder between T and Y , without capturing any post-treatment covariate. Hence, E' satisfies the strong ignorability: $Y(t) \perp\!\!\!\perp T | E'$.

5.2 Identifying Individualized Causal Effect

To identify the individualized causal effect, we assume that the variation of E' should be up to a piece-wise step function:

Assumption 5.8. (Piece-wise Variation) The Lebesgue measure of the set $\{\nabla_T f_\phi(\{Z_i\}_{i=1}^m, Y, T) \neq 0\}$ is zero.

A characteristic example of the above assumption is that the domain of E is discrete with finite values, or infinite values with zero measure (e.g., the set of rational numbers). With theorem 5.5 and the assumptions 5.6 and 5.8, we obtain the following identification theorem based on similar technique used in [27]:

Theorem 5.9. (Identification of the individualized causal effect) Assuming the consistency and piece-wise variation of the substitute E' , together with the separable assumption of outcome, we obtain a non-parametric identification on ICE by our method as follows:

$$\begin{aligned} & \mathbb{E}_Y[Y(t) | X] - \mathbb{E}_Y[Y(t') | X] \\ &= \mathbb{E}_{E'}[\mathbb{E}_Y[Y | X, T = t, E']] - \mathbb{E}_{E'}[\mathbb{E}_Y[Y | X, T = t', E']]. \end{aligned} \quad (12)$$

6 EXPERIMENTS

In this section, we empirically validate our method, CVAE-IV, in three IV prediction tasks with invalid IV violating the unconfounded assumption. Our experimental settings cover biased demand simulation with low and high dimensional features, and a realistic Mendelian randomization analysis. All the experiments are conducted in Python and Pytorch [21] framework.

Evaluation Metric We evaluate the root-mean-square error (RMSE) between the empirical ICE and the true ICE [10] to compare each method for verification as a continuous version of PEHE [24]:

$$R_{ICE} = \sqrt{\frac{1}{n} \sum_{k=1}^n ((Y_k(t) - Y_k(t')) - (\hat{Y}_k(t) - \hat{Y}_k(t')))^2}, \quad (13)$$

where $\hat{Y}_k(t')$ refers to the predicted counterfactual outcome with intervention as $T = t'$. The lower the calculated value of R_{ICE} , the better the obtained estimate for ICE is. Unlike previous IV methods which directly measure the divergence between g'_1 and g_1 [2, 6], our R_{ICE} instead measures the variation of the ICE.

Methods for Comparison We compare our proposed CVAE-IV with representative methods listed as follows: (a) Supervised method including DirectNN, as a feed-forward deep neural network (DNN) that directly predicts outcome without using IVs; (b) Two-stage IV methods, including 2SLS-Ploy as the two-stage least squares with polynomial basis functions [1], KernelIV as a kernelized method of two-stage regression strategy [25] and DeepIV as a deep two-stage regression method that employs a mixture density estimation technique [6]; (c) Proxy variable method including CEVAE, which models the joint distribution of all observed variables using a deep variational autoencoder [17]; (d) Invalid IV method including ModelIV, which ensemble predictions with each IV candidate based on modal validity [7].

Parameter Setup As shown in Fig. 2, our method comprises two main components: a conditional variational autoencoder for generation and a feed-forward neural network for regression. In more detail, we employ three hidden layers 128, 64 and 32 units respectively and ReLU activation functions for the encoder and the regression network [6]; moreover, the decoder is designed in three layers with [32, 64, 128] units and the same activation function. In addition, to deal with the high-dimensional airline demand analysis, we follow [6] by adopting a convolutional neural network (CNN), which has two convolutional layers with $64 \times 3 \times 3$ kernels, one maxpooling layer with pooling size 2×2 and two fully connected layers with 128 and 64 units, respectively. Our hyper-parameter λ for variance control in (6) is set as 0.1 across every setting. For DeepIV, CEVAE and KernelIV, we exactly follow the parameter settings provided in their original papers [6, 17, 25] with the same network architectures as stated above. For ModelIV, we also follow the original modal validity-based framework, where each estimate is learned via an independent DeepIV model [7]. For the three experiments, we set $n = 5000$ to facilitate training of both shallow and deep methods. Notably, we generate discrete confounders throughout our experiments to obey the piece-wise variation assumption.

6.1 Low-Dimensional Demand Simulation

Data Generation By modifying the widely used case [6, 7, 25], we simulate the airline demand analysis to include confounded IVs, which is illustrated in Fig. 1. Specifically, the observed covariates X include the customer types $s \sim \text{Unif}(\{1, 2, 3, 4, 5, 6, 7\})$ and the time of year $w \sim \text{Unif}(0, 10)$, where s exhibits different levels of price sensitivity and w models the effect of the holidays on both the sales Y and price T through a complex non-linear function $\psi_w = 2((w - 5)^4/600 + \exp[-4(w - 5)^2] + w/10 - 2)$. To include the invalid IVs that violates the unconfounded assumption, we consider three types of correlation between confounded IV candidate Z and unmeasured confounder of E :

[a] Z_{ef} is the direct effect of E :

$$Z_{ef} \sim \mathcal{N}(\beta_{ez}E, 1 - \beta_{ez}^2), \quad (14)$$

where $\beta_{ze} \in \mathcal{R}$ controls the correlation between Z_{ef} and E .

[b] Z_{ca} is the direct cause of E :

$$Z_{ca} \sim \mathcal{N}(0, 1) \quad \beta_{ze} \sim \text{Unif}(0, 1), \quad (15)$$

with $\beta_{ze} \in \mathcal{R}$ representing the correlation between Z_{ca} and E ;

[c] Z_{in} has indirect correlation with E :

$$V \sim \mathcal{N}(0, 1) \quad \beta_{ve}, \beta_{vz} \sim \text{Unif}(0, 1) \quad Z_{in} \sim \mathcal{N}(\beta_{vz}V, 1 - \beta_{vz}^2), \quad (16)$$

with $\beta_{ve}, \beta_{vz} \in \mathcal{R}$ as the coefficients controlling the correlations among variables. Notably, $E \sim \text{Bin}(\mathcal{N}(Z_{ca}^T \beta_{ze}^{ca} + \beta_{ve}V, 1), 100)$, where $\text{Bin}(\mathbf{K}, l)$ refers to discretizing vector \mathbf{K} into l possible bins which equally divides the

interval $[\min(\mathbf{K}), \max(\mathbf{K})]$. We then generate three scenarios by setting the number of IV candidates $m = 5$ and the sample size $n = 5000$:

- S1 Every IV candidate in $\{Z_i\}_{i=1}^5$ is a direct effect of E ;
- S2 Three IV candidates in $\{Z_1, Z_2, Z_3\}$ are direct causes of E with the other two as direct effects of E ;
- S3 Two IV candidates in $\{Z_1, Z_2\}$ are direct causes of E , one IV candidate Z_3 is indirectly correlated with E , one IV candidate Z_4 is a direct effect of E with the other valid IV Z_5 .

For each scene, the treatment and the outcome are generated via the following equations:

$$\begin{aligned} \beta_{zt} &\sim \text{Unif}(1.5, 2.5), \quad T = 25 + (\beta_{zt}^T \bar{Z} + 3)\psi_w + E, \\ Y &= 100 + (10 + r(T))s\psi_w - 2r(T) + E \end{aligned} \quad (17)$$

where \bar{Z} refers to the joint of $\{Z_i\}_{i=1}^5$ and $\beta_{zt} \in \mathcal{R}^5$ controls for correlation between IV candidates and treatment. Aside from the original linear setting which $r(T) = T$ [6], we adapt two non-linear settings [2], namely $r(T) = |T|$ and $r(T) = T^2$, to perform validation. For the former linear setting, we normalize T and Y with the same coefficients $\{T_\mu, T_{std}\}$ and $\{Y_\mu, Y_{std}\}$ used in [6], where we conduct normalization on T and Y in the latter two settings to force them to follow a normal distribution $\mathcal{N}(0, 1)$.

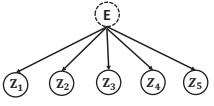
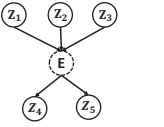
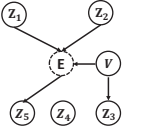
Analysis For airline demand simulation with low-dimensional features, we report the R_{ICE} for every method in Table 1. As is evident from the table, the trivial supervised method, DirectNN, obtains biased estimates under the influence of unmeasured confounder E . However, the classical IV prediction methods performs even worse than directly regression, which is caused by the incorrect usage of invalid IVs. As the violation of the unconfounded assumption entails the dependence between some IV candidates Z_i and E , applying two-stage regression methods in such cases obtains biased estimate T' that has a much stronger correlation with E than the original T in the treatment estimation stage, which further propagates error into regression stage and results in a far less accurate estimate of ICE [6]. For the proxy variable method CEVAE in our setting, its basic factorization of joint distribution is violated: the observed variable no longer independent with the treatment [17], resulting its biased estimation on ICE. Analogously, since the ModelIV [7] method is designed with a focus on violation of exclusion, ensembling single IV predictions fails to eliminate the bias brought by E . In contrary to the above methods, our method, namely CVAE-IV, achieves accurate estimation of R_{ICE} across three confounded IV settings by successfully recovering a reliable substitute E' of unmeasured confounder for identification of ICE.

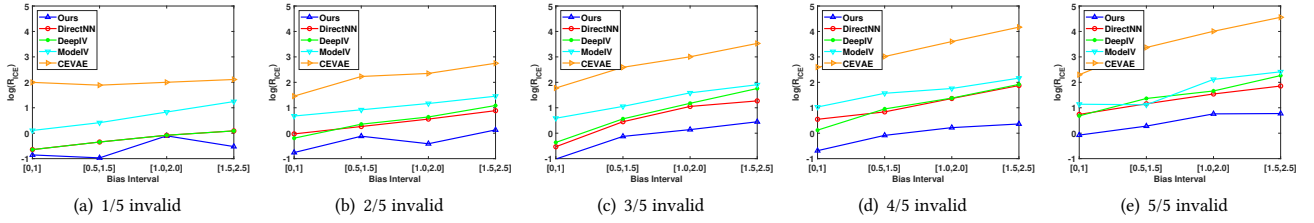
6.2 High-Dimensional Feature Space

Data Generation Following the previous protocol outlined in [6], we aim to test our method under a more challenging airline demand simulation with high-dimensional feature space. More specifically, the above mentioned customer type $s = \text{Unif}(\{1, 2, 3, 4, 5, 6, 7\})$ represents the individualized characteristics, while most realistic cases will not fall into such a uniformly delineated classes. Therefore, we relate each type number s to the pixel-wise feature that corresponds to that number in the MNIST dataset [6]. These high-dimensional features (784 dims) are directly used for individualized causal effect estimation, while the other components of the task remain exactly the same. For methods without deep architectures including 2SLS-Ploy and KernelIV, we feed them with 20-dimensional MNIST features obtained from a pretrained CNN [2].

Analysis As shown in Table 1, our method significantly outperforms other approaches with the high-dimensional features that R_{ICE} for each method is almost entirely consistent with those in the low-dimensional domain. Moreover, we analyze the performance on scenario 1 with high-dimensional features by simultaneously varying the number of invalid IVs (e.g., 1/5, 2/5...etc.) and the correlation strength β_{ze} between IV candidates and confounder. As shown in Fig. 3, when the correlation between Z_i and E

Table 1: Airline demand simulation in three scenarios with both low-dimensional and high-dimensional features: R_{ICE} averaged across ten runs, where the intervention $T = t'$ is set as a fixed grid of original price values t [6].

Scenario	Function	Dim	DirectNN	2SLS-Ploy	KernelIV	DeepIV	CEVAE	ModelIV	Ours
S_1 : 	Linear	Low	2.091	55.437	8.565	1.977	9.959	3.120	0.928
		High	2.069	56.476	9.220	2.261	10.283	4.426	0.858
	Abs	Low	1.874	44.144	6.893	2.426	8.481	1.988	0.697
		High	1.671	41.731	7.998	1.675	9.921	2.088	0.918
	Square	Low	1.414	52.872	8.521	1.163	6.569	2.423	0.490
		High	1.602	41.731	9.988	1.545	7.039	2.018	0.791
S_2 : 	Linear	Low	1.788	45.287	8.922	1.910	7.621	3.314	0.706
		High	2.152	56.484	9.215	1.788	8.109	2.040	0.601
	Abs	Low	1.595	34.214	5.864	1.224	7.178	2.064	0.562
		High	0.836	42.181	4.030	0.835	8.015	2.134	0.301
	Square	Low	1.168	41.108	5.749	1.064	10.385	1.565	0.125
		High	1.650	51.129	6.030	1.568	9.015	2.031	0.257
S_3 : 	Linear	Low	1.650	41.009	6.617	2.023	7.234	3.638	0.485
		High	1.821	41.374	7.656	1.729	7.203	4.134	0.608
	Abs	Low	2.095	44.471	4.867	1.148	9.789	1.916	0.689
		High	1.590	41.132	6.330	1.484	7.293	1.972	0.504
	Square	Low	1.111	41.199	7.826	1.019	8.362	1.580	0.516
		High	2.179	41.915	6.407	1.442	8.825	1.709	0.941

**Figure 3: R_{ICE} averaged across ten runs in the airline demand simulation with various numbers of invalid instruments, where the scenario in use is the first one and the effect function r is set to linear. The x-axis represents for the bias interval, which uniformly generates the correlation coefficients β_{ze} between confounder E and instruments Z_i .****Table 2: Performance averaged across ten runs on the Mendelian randomization analysis by varying p_{iv} from 20% to 100%, where we report the individualized causal effect of every method as R_{ICE} in this table.**

Methods	20% Invalid	30% Invalid	40% Invalid	50% Invalid	60% Invalid	70% Invalid	80% Invalid	100% Invalid
DirectNN	1.019	1.167	1.578	1.322	1.360	1.435	1.434	1.792
2SLS-Ploy	0.976	0.888	1.700	1.980	2.583	2.583	3.039	3.877
KernelIV	5.150	4.847	6.146	7.963	6.737	5.764	6.935	7.647
DeepIV	0.920	0.850	0.999	0.806	1.076	1.236	1.521	1.645
CEVAE	2.645	2.686	3.141	3.404	3.276	3.225	3.837	4.047
ModelIV	1.355	1.272	1.227	1.118	1.250	1.271	1.370	1.708
Our-method	0.276	0.428	0.459	0.304	0.566	0.376	0.846	0.717

is fixed, the bias of R_{ICE} for most methods (e.g., DirectNN, DeepIV, CEVAE) increases when we enlarge the number of invalid IV candidates. Comparatively, the proposed CVAE-IV remains insensitive to the variation of β_{ze} or proportion of invalid IVs.

6.3 Mendelian Randomization Analysis

Finally, we evaluate our method by adapting from a well known experimental analysis: Mendelian randomization studies [7] to reflect the violation of unconfounded assumption of IV prediction. Similarly to the demand

simulation case, we generate the IV candidates into four classes, with the proportion of invalid IVs denoted as p_{iv} :

[a] Valid IV candidates: $Z_{va} = \{Z_i \sim \text{Binomial}(2, q(i)) \mid i = 1, 2, \dots, [(1 - p_{iv}) * m]\}$ representing SNPs as locations in the genetic sequence, where $q(i) \sim \text{Unif}(0.1, 0.9)$ denotes the frequency with which an individual gets one or both rare genetic variants [7];

[b] Invalid IV candidates that directly cause E : Z_{ca} with each $Z_{ca} \in Z_{ca}$ generated exactly as in (15) and $|Z_{ca}| = \lfloor 0.3p_{iv} * m \rfloor$;

[c] Invalid IV candidates that indirectly correlate with E : Z_{in} with each $Z_{in} \in Z_{in}$ generated exactly as in (16) and $|Z_{in}| = \lfloor 0.3p_{iv} * m \rfloor$;

[d] Invalid IV candidates that are directly effected by E: Z_{ef} with each $Z_{ef} \in Z_{ef}$ generated exactly as in (14) and $|Z_{ef}| = \lfloor 0.3p_{iv} * m \rfloor$;

Then we generate $E = \text{Bin}(\mathcal{N}(E_{ca} + E_{in}, 1), 100)$, together with IV candidates as $\bar{Z} = [Z_{va}, Z_{ca}, Z_{in}, Z_{ef}]$ and observed features $X \in \mathcal{R}^{10}$ as $X(i) \sim \text{Unif}(-0.5, 0.5)$. To enable estimation of ICE, we generate treatment and outcome as follows:

$$T = \sum_{i=1}^m \alpha_i Z_i + E + \epsilon_x \quad \beta(X) = \text{round}\left(X^T \gamma^{(xt)}, 0.1\right) \quad (18)$$

$$Y = \beta(X)T + \sum_{i=1}^m \delta_i Z_i + E + \epsilon_y,$$

where α_i modulates genetic variable effect on the treatment is given by $\alpha_i = \frac{\sqrt{0.1}}{\sigma_{zx}} v_i$, $v_i \sim \text{unif}(0.01, 0.2)$ and $\sigma_{zx} = \text{std}(\sqrt{0.1} \sum_i v_i Z_i)$ [7]. Meanwhile, coefficient $\gamma^{(xt)}$ representing individualized causal effect is a sparse vector of length 10, with three non-zeros $\gamma^{(xt)}(i) \sim \text{Unif}(0.2, 0.5)$. The error terms ϵ_x and ϵ_y are independently generated from a normal distribution with mean 0. In this experiment, we perform validation by setting the number of IV candidates $m = 100$ in a much larger size. Following the previous experience [7], we also plug the inductive bias, namely the conditionally linearity in treatment effect function, into the parameterization of neural network: $g_{\phi_1}(X, T) = h_1(h(X))T + h_2(h(X))$, where h_1 and h_2 are linear layers shared on representation h .

Analysis As shown in Table 2, we report the general trends on Mendelian randomization experiment by varying proportion of invalid IVs p_{iv} . It is obvious that when p_{iv} arises, R_{ICE} of most methods also increases except for ours. Notably, 2SLS-Ploy performs than most IV methods with deep architectures, due to the conditional linear relationship between treatment and outcome. Moreover, we infer that most deep methods (e.g., CEVAE, DeepIV) tend to overfit on training data and thus obtain more seriously biased estimate than 2SLS. Comparatively, our method achieve stable and accurate across the variation of invalid proportion, which further verifies the effectiveness of CVAE-IV.

6.4 Ignorability Analysis

To provide a deeper insight into the behaviour of our method, we also test the ignorability of the generated substitute E' from two aspects: (a) If $E \perp\!\!\!\perp T \mid E', X$, then ignorability obviously holds due to the separable formulation in the IV problem. (b) If there is a measurable function f such that $E = f(E')$, then ignorability holds, due to $f(E') \perp\!\!\!\perp T \mid E', X$ by the strong ignorability of E and $g_2(E) = E$ in our experiments. For the former criterion, we test the independence between E and T given the generated E' via a classifier-based conditional independence testing method named CCIT [23] (using p_value), while we construct a four-layer fully connected network with hidden unit numbers of [128, 63, 32] from E' to E to test the second criterion (using RMSE). As shown in Table 3, in the Mendelian randomization scenario with varying proportions of invalid IVs, both the fact that the p_value of our methods is uniformly distributed and the low regression error imply that the ignorability of E' holds. For the airline demand scenario, although the second criterion does not hold (E' might contain some noise variable independent from E), the conditional independence results still reflect that E' is ignorable.

7 LIMITATION AND FUTURE WORK

In this paper, we concentrate on estimating individualized causal effect with invalid IVs that violate the unconfounded assumption. By constructing the CVAE-IV model to generate a ignorable confounder substitute, we isolate the influence of the unmeasured confounder from the estimation on ICE. Experiments on airline demand simulation and Mendelian randomization analysis verify the validity of our method with nearly unbiased estimation results of ICE. However, we only consider the separable formulation

Table 3: Verifying ignorability on confounder substitute. Here, we choose the linear setting of airline demand simulation under scenario 1 and the Mendelian Randomization analysis with p_{iv} set to 20%, 50% and 100%, respectively.

Scenario	Feature	P_value	RMSE
Demand_Scenario_1_linear	low	0.290	0.210
Demand_Scenario_1_linear	high	0.316	0.102
Demand_Scenario_2_linear	low	0.886	0.310
Demand_Scenario_2_linear	high	0.966	0.250
Demand_Scenario_3_linear	low	0.506	0.130
Demand_Scenario_3_linear	high	0.456	0.150
Mendelian_20%	low	0.956	0.013
Mendelian_50%	low	0.766	0.054
Mendelian_100%	low	0.238	0.008

in the IV problem [6]. Solutions on non-separable cases will involve considering the control function variable by building the connection between treatment-IV and treatment-confounder correlations [22].

7.1 Acknowledgement

This work was supported in part by National Natural Science Foundation of China (No.91948303-1, No. 62006207, No. 61803375, No. 62101575, No. 61906210, No. 62037001), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Key Laboratory for Corneal Diseases Research of Zhejiang Province, Project by Shanghai AI Laboratory (P22KS 00111) and the Fundamental Research Funds for the Central Universities (226-2022-00142). We would also like to thank Dr. Adi Lin for productive discussions.

REFERENCES

- [1] Takeshi Amemiya. 1974. The nonlinear two-stage least-squares estimator. *Journal of econometrics* 2, 2 (1974), 105–110.
- [2] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep Generalized Method of Moments for Instrumental Variable Analysis. *NeurIPS* 32 (2019), 3564–3574.
- [3] Debo Cheng, Jiuyong Li, Lin Liu, Jiji Zhang, Jixue Liu, et al. 2022. Ancestral instrument method for causal inference without a causal graph. *arXiv preprint arXiv:2201.03810* (2022).
- [4] Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill DF Campbell, and Ivor Simpson. 2018. Structured uncertainty prediction networks. In *CVPR*. 5477–5485.
- [5] Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. 2018. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *JRSSB* 80, 4 (2018), 793–815.
- [6] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *ICML*. PMLR, 1414–1423.
- [7] Jason S Hartford, Victor Veitch, Dhanya Sridhar, and Kevin Leyton-Brown. 2021. Valid causal inference with invalid instruments. In *ICML*. PMLR, 4096–4106.
- [8] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. 2017. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology* 46, 6 (2017), 1985–1998.
- [9] Gibran Hemani, Jack Bowden, and George Davey Smith. 2018. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* 27, R2 (2018), R195–R208.
- [10] Guido W Imbens and Whitney K Newey. 2009. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77, 5 (2009), 1481–1512.
- [11] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- [12] Olav Kallenberg and Olav Kallenberg. 1997. *Foundations of modern probability*. Vol. 2. Springer.
- [13] 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. *JASA* 111, 513 (2016), 132–144.
- [14] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *SIGKDD*. 265–274.

- [15] 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. In *AISTATS*. PMLR, 398–410.
- [16] Sai Li and Zijian Guo. 2020. Causal inference for nonlinear outcome models with possibly invalid instrumental variables. *arXiv preprint arXiv:2010.09922* (2020).
- [17] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *NeurIPS*. 6449–6459.
- [18] Marloes H Maathuis and Diego Colombo. 2015. A generalized back-door criterion. *The Annals of Statistics* 43, 3 (2015), 1060–1088.
- [19] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105, 4 (2018), 987–993.
- [20] Krikamol Muandet, Arash Mehrjou, Si Le Kai, and Anant Raj. 2020. Dual Instrumental Variable Regression. In *NeurIPS 2020*.
- [21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* 32 (2019).
- [22] Aahlad Puli and Rajesh Ranganath. 2020. General Control Functions for Causal Effect Estimation from Instrumental Variables. *NeurIPS* 33 (2020), 8440.
- [23] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. 2017. Model-Powered Conditional Independence Test. *NeurIPS* 30 (2017).
- [24] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. PMLR, 3076–3085.
- [25] Rahul Singh, Maneesh Sahani, and Arthur Gretton. 2019. Kernel instrumental variable regression. *arXiv preprint arXiv:1906.00232* (2019).
- [26] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *NeurIPS* 28 (2015), 3483–3491.
- [27] Yixin Wang and David M Blei. 2019. The blessings of multiple causes: rejoinder. *JASA* 114, 528 (2019), 1616–1619.

8 APPENDIX

8.1 Preliminaries

We introduce two lemmas from [12] for further construction of our theory.

Definition 8.1. (Probability Kernel) Given two measurable spaces (S, \mathcal{S}) and (T, \mathcal{T}) , a mapping $\mu : S \times T \rightarrow \mathcal{R}_+$ is called a probability kernel from S to T if the function $\mu_s(B) = \mu(s, B)$ is \mathcal{S} -measurable in $s \in S$ for fixed $B \in \mathcal{T}$ and a probability measure in $B \in \mathcal{T}$ for fixed $s \in S$.

Lemma 8.2. (Conditional distribution) Fix a Borel space S and a measurable space T , and let ξ and η be random elements in S and T , respectively. Then there exists a probability kernel μ from T to S satisfying $P[\xi \in \cdot | \eta] = \mu(\eta, \cdot)$ almost surely.

Lemma 8.3. (Kernels and randomization) Let μ be a probability kernel from a measurable space S to a Borel space T . Then there exists some measurable functions $f : S \times [0, 1] \rightarrow T$ such that if $\mathcal{V} \sim U(0, 1)$, then $f(s, \mathcal{V})$ follows distribution $\mu(s, \cdot)$ for every $s \in S$.

8.2 Ignorability of Confounder Substitute

Assumption 8.4. (Weak regularity condition.) Consider a random variable Z from basic space (Ω, \mathcal{F}, P) to its domain \mathcal{Z} , Z satisfies weak regularity condition iff \mathcal{Z} is a Borel space.

Theorem 8.5. Every CVAE Model $(\{Z\}_{i=1}^m, \{\theta_z, \theta_y\}, E', T)$ admits a Kallenberg construction under weak regularity condition.

PROOF. First, without loss of generality, we assume that the product space of each domain $\mathcal{Z}^m = \times_{j=1}^m \mathcal{Z}_j$ is a Borel subset of compact set, which can be easily reduced to case that $\mathcal{Z}^m = [0, 1]^m$ [12].

Then, due to the fact that $\bar{Z} = (Z_1, Z_2, \dots, Z_m)$ and (E', T) are random variables in \mathcal{Z}^m and $\mathcal{E}' \times \mathcal{T}$, respectively, Lemma 8.2 guarantee that there exists a probability kernel that reflects the conditional distribution $P(\bar{Z} | E', T)$. Combined with the fact we could easily find a random variable z in $[0, 1]$ such that $z \perp (E', T)$, Lemma 8.3 further implies the existence of measurable function $f : (\mathcal{E}' \times \mathcal{T}) \times [0, 1] \rightarrow \mathcal{Z}^m$ such that:

$$\bar{Z} = f((E', T), z). \quad (19)$$

Similar to the above derivation, there exists measurable function $h_z : \Theta_z \times [0, 1] \rightarrow [0, 1]$ and random variables W_z in $[0, 1]$ such that:

$$z = h_1(\theta_z, W_z), \quad (20)$$

where we could easily let $W_z \sim U(0, 1)$ such that $W_z \perp (E', T, \theta_z)$ since θ_z are point masses and their σ -algebra is trivial.

In an analogous approach, we construct a measurable function $g : \mathcal{E}' \times [0, 1] \rightarrow \mathcal{Y}$ such that there exists a variable $\gamma_y \perp (E', T)$ such that:

$$Y = g((E', T), \gamma_y), \quad (21)$$

where the γ_y is induced from a measurable function $h_y : \Theta_y \times [0, 1] \rightarrow [0, 1]$ with a random variable $W_y \sim U(0, 1)$ and $W_y \perp (E, T, \theta_y)$ such that:

$$\gamma_y = h_1(\theta_y, W_y), \quad (22)$$

Then we argue that the set of variables $\{W_z, W_y\}$ are independent, due to the fact that $\bar{Z} = f((E', T), h_z(\theta_z, W_z))$ and $Y = g((E', T), h_y(\theta_y, W_y))$ are independent conditioning on E', T . (Here we omit θ in condition due to the fact that $\theta_{y,z}$ are point masses and their trivial σ -algebra renders it naturally independent of any other variable).

Finally, the inverse trick $U = F^{-1}(y)$ and $V = G^{-1}(z)$ renders that $U, V \sim U(0, 1)$, and the following equation further holds:

$$\bar{Z} = f((E', T), U), \quad (23)$$

and

$$Y = g((E', T), V), \quad (24)$$

where F and G is the cumulative distribution function of z and y . Observe that $U = F^{-1}(h_z(\theta_z, W_z))$ and $V = G^{-1}(h_y(\theta_y, W_y))$, it is obvious that $\sigma(U) \subseteq (\sigma(\theta_z) \cap \sigma(W_z))$ and the same holds for V , which further implies that $\sigma(U)$ and $\sigma(V)$ are independent. \square

Theorem 8.6. Any hidden confounder E must be captured by the learned CVAE model through the generated substitute E' .

PROOF. First, we refer that variable A is captured by variable B iff the generative σ -algebra of A is contained in that of B . Then we start our proof by contradiction. Suppose there exists an unobserved confounder E_0 that is not captured by the generated E' , that is, $\sigma(E_0)$ cannot be measured with respect to $\sigma(E')$, then Lemma 8.3 implies that there exists functions f and g such that $(Z_1, Z_2, \dots, Z_m) = f((E', T), U)$ and $Y = g((E', T), V)$ holds. Moreover, the definition of hidden confounder E_0 renders that $U = I_1(E_0, \gamma_1)$ ($\gamma_1 \perp E_0$). Meanwhile, due to the fact that there always exists correlation between \bar{Z} and E_0 once we conditioning on T (either the correlation between Z_j and E_0 for some j or the correlation between \bar{Z} and E_0 conditioning on T for every j), we could write $V = I_2(E_0, \gamma_2)$ ($\gamma_2 \perp E_0$) without loss of generality. Now turn to the fact that $U = I_1(E_0, \gamma_1)$ ($\gamma_1 \perp E_0$) and $V = I_2(E_0, \gamma_2)$, we obtain that $U \not\perp V$, because that E_0 is not measurable with respect to $\sigma(E')$. This conclusion contradicts with the conditional Kallenberg construction principle conveyed in our CVAE model that $U \perp V$. \square

Assumption 8.7. (Consistency) The estimation of CVAE model is consistent if and only if for some function f_ϕ ,

$$p(E | \bar{Z}, Y, T) = \delta(f_\phi(\bar{Z}, Y, T)) \quad (25)$$

Theorem 8.8. Any post-treatment variable V of T must not be measured with generated confounder substitute E' .

PROOF. We prove this conclusion by contradiction. Without loss of generality, we assume that E' picks up a post-treatment variable V and separate E' into E'_V and E'_{other} , where E'_V is the post-treatment component and E'_{other} is the other component. Then the consistency condition implies that:

$$p(E' | T, \bar{Z}, Y) = \delta(f_v(T, \bar{Z}, Y; \phi), f_o(T, \bar{Z}, Y; \phi)), \quad (26)$$

where f_v and f_o are deterministic functions that mapping T, \bar{Z}, Y to E'_V and E'_{other} . Meanwhile, the conditional distribution of E'_V given T, \bar{Z}, Y is:

$$\begin{aligned} p(E'_V | T, \bar{Z}, Y) &= p(E'_V | T, \bar{Z}, Y, E'_{\text{other}}) \\ &= \frac{p(E'_V | E'_{\text{other}}, \bar{Z}, Y) p(T | E'_V, E'_{\text{other}}, \bar{Z}, Y)}{p(T | E'_{\text{other}}, \bar{Z}, Y)}, \end{aligned} \quad (27)$$

where the first equality is due to the consistency condition. Combining (26) and (27), we conclude that either $p(E'_V | E'_{\text{other}}, \bar{Z}, Y)$ or $p(T | E'_V, E'_{\text{other}}, \bar{Z}, Y)$ is a point mass, which both entails that the learned CVAE model is “degenerate”, i.e., non-probabilistic [27]. More specifically, the former implies that some component E'_V in latent representation of CVAE is a deterministic function of other components E'_{other} , where the latter implies that the input T is a deterministic function of latent representation E' . Thus the contradiction happens with the fact that CVAE is a deep probabilistic model. \square

8.3 Identification of Individualized Causal Effect

Assumption 8.9. (Piece-wise Variation) The Lebesgue measure of the set $\{\nabla_{\mathbf{T}} f_{\phi}(\{\mathbf{Z}_i\}_{i=1}^m, \mathbf{Y}, \mathbf{T}) \neq 0\}$ is zero.

With theorem 8.6 and the assumptions 8.7 and 8.9, we obtain the following identification theorem based on similar technique in [27]:

Theorem 8.10. (Identification of the individualized causal effect) Assume consistency and piece-wise variation of the substitute \mathbf{E} , together with the separable assumption of outcome, we obtain a non-parametric identification on individualized causal effect by our method as follows:

$$\begin{aligned} & \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{X}] - \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t}')|\mathbf{X}] \\ &= \mathbb{E}_{\mathbf{E}'}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}, \mathbf{E}']] - \mathbb{E}_{\mathbf{E}'}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}', \mathbf{E}']], \end{aligned} \quad (28)$$

PROOF. First, we expand the L.H.S in equation (28) as follows:

$$\begin{aligned} & \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{X}] - \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t}')|\mathbf{X}] \\ &= \mathbb{E}_{(\mathbf{E}|\mathbf{X})}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{E}, \mathbf{X}]] - \mathbb{E}_{(\mathbf{E}|\mathbf{X})}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t}')|\mathbf{E}, \mathbf{X}]] \\ &= ([g_1(\mathbf{t}, \mathbf{X})] + \mathbb{E}_{\mathbf{E}|\mathbf{X}}[g_2(\mathbf{E})]) - ([g_1(\mathbf{t}', \mathbf{X})] + \mathbb{E}_{\mathbf{E}|\mathbf{X}}[g_2(\mathbf{E})]) \\ &= g_1(\mathbf{t}, \mathbf{X}) - g_1(\mathbf{t}', \mathbf{X}) \\ &= \int_{\mathbf{t}}^{\mathbf{t}'} \nabla_{\mathbf{T}} g_1(\mathbf{T}, \mathbf{X}) d\mathbf{T}, \end{aligned} \quad (29)$$

where the first equation is due to tower property, the second equation is due to the separable formulation of potential outcome in (1), and the last equation is due to fundamental theorem of Calculus (Here we assume $\mathbf{t} \leq \mathbf{t}'$ without loss of generality). Meanwhile, we derive some results with respect to the R.H.S of (28) as follows:

$$\begin{aligned} & \nabla_{\mathbf{T}} \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}, \mathbf{E}' = \mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})] \\ &= \nabla_{\mathbf{T}} \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{X}, \mathbf{T} = \mathbf{t}, \mathbf{E}' = \mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})] \\ &= \nabla_{\mathbf{T}} \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{X}, \mathbf{E}' = \mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})] \\ &= \nabla_{\mathbf{T}} g_1(\mathbf{t}, \mathbf{X}) + \nabla_{\mathbf{T}} g_2(\mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})) \\ &= \nabla_{\mathbf{T}} g_1(\mathbf{t}, \mathbf{X}) + \nabla_{f_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})} g_2 * \nabla_{\mathbf{T}} (f_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})) \\ &\stackrel{a.e.}{=} \nabla_{\mathbf{T}} g_1(\mathbf{t}, \mathbf{X}), \end{aligned} \quad (30)$$

where the first equality is due to SUTVA [11], the second equality is due to the ignorability by \mathbf{E}' in Theorem 8.6, the third equality is due to separable formulation, the fourth equality is due to chain rule and the last equality is due to the piece-wise variation assumption. Then we obtain an important fact for our identification based on separable formulation of the outcome regression model:

$$\begin{aligned} & \nabla_{\mathbf{T}} \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}, \mathbf{E}' = \mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})] \\ &= \nabla_{\mathbf{T}} g_{\psi_1}(\mathbf{X}, \mathbf{t}) + \nabla_{\mathbf{T}} g_{\psi_2}(\mathbf{f}_{\phi}(\mathbf{Y}, \bar{\mathbf{Z}}, \mathbf{T})) \\ &\stackrel{a.e.}{=} \nabla_{\mathbf{T}} g_{\psi_1}(\mathbf{t}, \mathbf{X}), \end{aligned} \quad (31)$$

where the second equality is due to the chain rule and piece-wise variation assumption. Combining the equation (30) and (31), we bridge the R.H.S and L.H.S of (28) as follows:

$$\begin{aligned} & \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t})|\mathbf{X}] - \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(\mathbf{t}')|\mathbf{X}] \\ &= \int_{\mathbf{t}}^{\mathbf{t}'} \nabla_{\mathbf{T}} g_1(\mathbf{T}, \mathbf{X}) d\mathbf{T} \\ &= \int_{\mathbf{t}}^{\mathbf{t}'} \nabla_{\mathbf{T}} g_{\psi_1}(\mathbf{T}, \mathbf{X}) d\mathbf{T} \\ &= g_{\psi_1}(\mathbf{t}, \mathbf{X}) - g_{\psi_1}(\mathbf{t}', \mathbf{X}) \\ &= (g_{\psi_1}(\mathbf{t}, \mathbf{X}) + \mathbb{E}_{\mathbf{E}'}[g_{\psi_2}(\mathbf{E}')] - (g_{\psi_1}(\mathbf{t}', \mathbf{X}) + \mathbb{E}_{\mathbf{E}'}[g_{\psi_2}(\mathbf{E}'])) \\ &= \mathbb{E}_{\mathbf{E}'}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}, \mathbf{E}']] - \mathbb{E}_{\mathbf{E}'}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}, \mathbf{T} = \mathbf{t}', \mathbf{E}']]. \end{aligned} \quad (32)$$

where the fourth equality is due to fundamental theorem of calculus, the fifth equality is due to separable formulation of outcome model, the last equality is due to definition of expectation. \square