Distribution-Conditioned Adversarial Variational Autoencoder for Valid Instrumental Variable Generation

Xinshu Li[1*], **Lina Yao**[1, 2]

[1] School of Computer Science and Engineering, The University of New South Wales
[2] CSIRO's Data 61
xinshu.li@unsw.edu.au, lina.yao@data61.csiro.au

Abstract

Instrumental variables (IVs), widely applied in economics and healthcare, enable consistent counterfactual prediction in the presence of hidden confounding factors, effectively addressing endogeneity issues. The prevailing IV-based counterfactual prediction methods typically rely on the availability of valid IVs (satisfying **Relevance**, **Exclusivity**, and **Exogeneity**), a requirement which often proves elusive in real-world scenarios. Various data-driven techniques are being developed to create valid IVs (or representations of IVs) from a pool of IV candidates. However, most of these techniques still necessitate the inclusion of valid IVs within the set of candidates. This paper proposes a distribution-conditioned adversarial variational autoencoder to tackle this challenge. Specifically: 1) for **Relevance** and **Exclusivity**, we deduce the corresponding evidence lower bound following the Bayesian network structure and build the variational autoencoder accordingly; 2) for **Exogeneity**, we design an adversarial game to encourage latent factors originating from the marginal distribution, compelling the independence between IVs and other outcome-related factors. Extensive experimental results validate the effectiveness, stability and generality of our proposed model in generating valid IV factors in the absence of valid IV candidates.

Introduction

Counterfactual prediction has attracted increasing attention (Alaa and van der Schaar 2017; Li et al. 2016; Chernozhukov, Fernández-Val, and Melly 2013; Glass et al. 2013) in recent years due to the rising demands for robust and trustworthy artificial intelligence. *Confounders*, the common causes of treatments and effects, induce spurious relations between different variables, resulting in famous endogeneity issues in causal inference. Quantities of related work (Li and Yao 2022; Yao et al. 2018; Yoon, Jordon, and Van Der Schaar 2018; Shalit, Johansson, and Sontag 2017) try to mitigate the bias caused by confounders under the "*no hidden confounders*" assumption, which is untestable and impractical in the real world. Therefore, recent research interests have switched the focus to counterfactual prediction with unobserved confounders.

Instrumental variable (IV) plays a crucial role in addressing counterfactual prediction in the presence of unobserved confounders. The core concept behind IVs is to incorporate a powerful exogenous variable that correlates with the treatment but remains independent of the hidden confounders, thereby eliminating endogeneity issues and obtaining more accurate and consistent estimates. To ensure its effectiveness and applicability, a *valid* IV is supposed to satisfy three conditions: 1) **Relevance**: IV must be correlated with the treatment variable; 2) **Exclusivity**: there should be no direct causal relationship between the IV and the outcome variable; 3) **Exogeneity**: IV should have no correlation with the unobserved confounders to ensure their influence on the outcome variable is solely through the treatment variable.

For example, as shown in Figure 1, Hoxby (2000) investigated whether competition among public schools (Treatment T) enhances educational quality within districts (Outcome Y). The potential endogeneity of the regional school counts arises as it and the outcome might be simultaneously influenced by long-term factors (Confounders) specific to that area. Some confounders (observed Confounders C) can be quantified, such as the local economic development level. More confounding variables (Unobserved Confounders U), however, cannot be quantified and exhaustively enumerated, such as certain historical factors. River counts here can serve as a persuasive IV: 1) more rivers can lead to more schools due to transportation (Relevance); 2) yet they are unrelated to teaching quality directly (Exclusivity); 3) the natural attributes of river formation render the counts of the rivers unrelated to the confounders caused by historical factors (Exogeneity).

Two-stage least squares (2SLS) regression method (Angrist and Imbens 1995) is a typical IV-based algorithm for counterfactual prediction with hidden confounders under a linear setting. Recently, some works (Singh, Sahani, and Gretton 2019; Wu et al. 2022; Xu et al. 2020; Hartford et al. 2017; Lin et al. 2019) have introduced deep learning techniques into this area and outperformed classical 2SLS methods in the nonlinear scenario. However, the prerequisite for these methods is to access valid IVs, which remains challenging in real-world scenarios.

Extensive literature (Hartford et al. 2021; Yuan et al. 2022; Davies et al. 2015; Burgess, Dudbridge, and Thompson 2016) has attempted to resolve this problem by generat-

*Corresponding author.

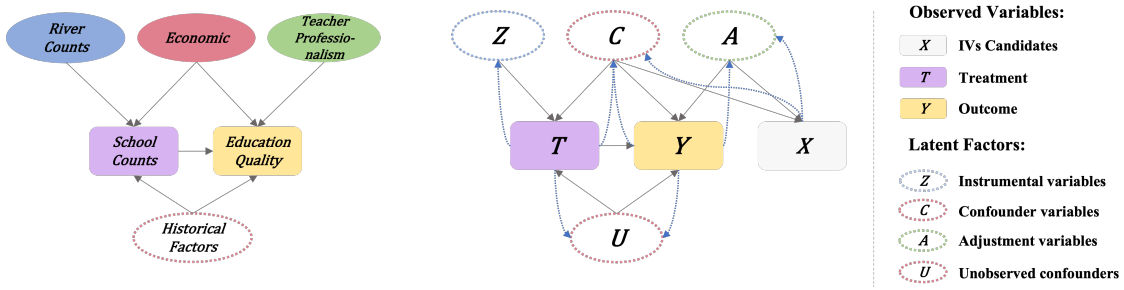

Figure 1: (Left) Causal graph of the proposed example (Hoxby 2000). School counts is the treatment T, educational quality is the outcome Y, and local economic status is the observed confounders C. While not directly measurable, some historical factors act as unobserved confounders U that impede counterfactual prediction as well. The river counts within the districts play a role as instrumental variables Z, which influence education quality only via school counts. A represents the factors that only have causal relations with outcomes, e.g., teacher professionalism. (Right) Bayesian network structure corresponds to the causal graph. Black arrows with solid lines denote the generative process, and blue arrows with dashed lines indicate the inference process. We study the more general scenario in real life, while IV candidates X is a view of C and A without Z.

ing IVs automatically from observed features, or rather, IVs candidates. However, most of these methods share a common drawback: valid strong IVs must exist in the candidate set for effective instrumental variable generation using different weighting or representation methods. Recently, GIV (Wu et al. 2023) models group variable as a valid IV in the absence of valid IVs candidate. However, this method shows effectiveness solely in scenarios where the dataset is derived from diverse sources, which restricts its application scope.

It has been a common practice to introduce latent factors to reflect all exogenous uncertainty (Hartford et al. 2017; Kim et al. 2021; Pfohl et al. 2019) in causal reference. Nevertheless, despite the focus on model confounders as latent factors, how to design deep latent models to generate IVs remains barren. In this paper, we develop a deep **V**ariational autoencoder to generate **I**nstrumental **V**ariables, **VIV**, in the latent space.[1] To guarantee the validness of the generated VIV, we: 1) model the exogenous uncertainty from multiple factors, either directly related to outcomes, e.g., (un)observed confounders, or indirectly related to outcomes, e.g., IVs for **Exclusivity**; 2) construct the inference and generative networks based on the Bayesian network structure for **Relevance**; 3) design an adversarial game encouraging the joint distribution of latent factors align with their marginal distributions for **Exogeneity**.

Our main contributions are summarized as follows:

- We address the issue of IV-based counterfactual prediction when there are no valid instrumental variables candidates available, a situation that has been difficult to navigate with previous methods of IV generation.
- We propose a novel distribution-conditioned adversarial variational autoencoder VIV for valid IV generation, which uniquely employs VAE's inference network as GAN's generative network. The VAE framework built upon a causal graph, along with adversarial learning to encourage latent factors originating from the marginal distribution, collectively ensures strict compliance with valid instrument conditions.
- We plug our generated IVs into the downstream prevailing IV-based counterfactual prediction algorithms. The extensive experimental results validate the superiority of our method in generating valid IVs in the lack of valid IV candidates compared with the state-of-the-art IV generation methods.

[1]For the sake of brevity, we also name our method as VIV.

Related Work

IV-based Counterfactual Prediction

In the linear setting, 2SLS (Angrist and Imbens 1995) is a classical method to apply IV in counterfactual prediction. In the non-linear scenario, Singh, Sahani, and Gretton (2019); Muandet et al. (2020) adopts kernel methods to capture non-linear relations between variables. Hartford et al. (2017); Lin et al. (2019); Xu et al. (2020) introduces deep learning techniques and fits a mixture density network. Bennett, Kallus, and Schnabel (2019); Dikkala et al. (2020) use moment conditions for model parameters estimation. Wooldridge (2015); Puli and Ranganath (2020) set up a control function estimator for causal inference with IVs. These approaches necessitate precisely predefined independent variables (IVs), a condition rarely met in practical scenarios. Consequently, their ability to effectively apply to real-world situations is inevitably compromised.

IV Generation

The challenge of identifying valid instrumental variables has led to a proliferation of research focused on generating instrumental variables from a vast pool of candidates, primarily derived from observed variables. Burgess, Dudbridge, and Thompson (2016); Kuang et al. (2020) weight the IVs candidates equally (UAS) or according to the correlation between them and treatments (WAS) to create summary IVs. Furthermore, Hartford et al. (2021) employs the closest cluster center of estimation points as an instrumental variable. Yuan et al. (2022) generates IV presentations

by constraining the information flow between different variables. These IV-generation methods necessitate a collection of high-quality IV candidates. However, this is unfeasible in real-world scenarios due to financial constraints and a need for specialized expertise. Wu et al. (2023) attempt to learn a group variable as the instrument to denote the source in the absence of valid IV candidates. Nevertheless, its effectiveness is confined to scenarios where data is gathered from multiple sources, limiting its broader range of applications.

Deep Latent Models in Causal Inference

Louizos et al. (2017) introduces variational autoencoder (Kingma and Welling 2013) into causal inference. They model the hidden confounder as the latent variable and assume that the observed feature is a noisy view of the hidden confounders. Pfohl et al. (2019); Kim et al. (2021) modify the causal graph of (Louizos et al. 2017) and propose more general latent models derived from the adjusted causal structure. Nonetheless, the focal point of these endeavors centers around capturing exogenous uncertainty to accomplish causal inference or causal fairness tasks. Consequently, they only assess exogenous uncertainty from the perspective of confounders. The exploration of how to establish deep latent variable models for the generation of instrumental variables has remained a gap in the research. Compared with previous deep latent models in causal inference, VIV is the early pioneer of measuring exogenous uncertainty from a broader perspective and generating instruments for downstream counterfactual prediction.

Preliminaries

As shown in Figure 1, we have observed pre-treatment features X, treatment T, and outcome Y in the dataset $\mathcal{D}$. IV-generation methods commonly refer to X as the pool of instrument variables candidates, as the synthesized IVs stem from weighted or latent representations of X. From the perspective of the Data Generating Process (**DGP** (Zimmermann et al. 2021)), the observed variables stem from multiple exogenous latent factors. This paper centers on three types of such factors: **Instrumental variable** Z that only directly influences T; **Confounder variable** that directly affects both T and Y; **Adjustable variable** A that only directly influences Y.

Different from previous IV-generation work (Kuang et al. 2020; Hartford et al. 2021; Yuan et al. 2022), which assume X is a view of Z, **observed Confounders** C and A (optional), we challenge a more demanding scenario where Z cannot be captured by X. Besides C, some **Unobserved confounders** U are not reflected in X but impede the counterfactual prediction.

Given the above definitions, we present the formal definitions of valid instrumental variables and counterfactual prediction problem, followed by an essential assumption for identification adopted in this paper.

Definition 0.1. ***Valid Instrumental Variables*** *Z are characterized by satisfying the subsequent conditions (Hartford et al. 2017):*

Relevance*: Z must be correlated with the treatments T, i.e., $P(T|Z) \neq P(T)$.*

Exclusivity*: Z cannot be a direct causal parent of the outcomes Y, i.e., $P(Y|Z,T,A,C,U) = P(Y|T,A,C,U)$.*

Exogeneity*: IV is independent with the confounders, i.e., $P(Z|C,U) = P(Z)$.*

Definition 0.2. ***Counterfactual Prediction*** *refers to predicting what the outcome would have been for an individual under an intervention t (Pearl et al. 2000), i.e.,*

$$g(t,X) = E[Y|do(T=t),X]. \tag{1}$$

Assumption 0.3. ***Additive Noise Assumption:*** *the noise form unmeasured confounders U is added to the outcomes Y (Hartford et al. 2017), i.e.,*

$$Y = g(T,X) + U. \tag{2}$$

Our strategy for generating valid instruments is to utilize variational autoencoder (Kingma and Welling 2013; Higgins et al. 2017; Dupont 2018; Kim and Mnih 2018; Wu and Fukumizu 2022) to deduce the intricate nonlinear connections between latent factors Z, U, C, A and observable variables X, T, Y, thereby facilitating an approximate reconstruction of the joint distribution $p(X,T,Y,U,Z,C,A)$. A limitation associated with VAE-based models arises from the challenge of attaining global optima during the optimization of neural networks. Consequently, there is no guarantee that a specific instance, even within the model class, will converge to the true model. Despite this, we hold the belief that this disadvantage is counterbalanced by the robust empirical efficacy exhibited by deep neural networks across numerous domains (Fortuin et al. 2020; Razavi, Van den Oord, and Vinyals 2019; Kim et al. 2021).

Methodology

Identification of Causal Effects

This paper aims to achieve counterfactual prediction with the generated IVs, (un)observed confounders and adjustable variables. We first prove the identification in our case, which enables causal effects can be uniquely estimated from the observed data. It stems directly from Pearl's back-door adjustment formula (Pearl et al. 2000):

Theorem 0.4. *If we recover $p(X,T,Y,U,Z,C,A)$, then the causal effects under the causal structure in Figure 1 is identifiable.*

Proof. We will prove that $p(Y|do(T=t),X)$ is identifiable under the premise of 0.4. We have that:

$$\begin{aligned} & p(Y|do(T=t),X) \\ = & \iint_{U,C} p(Y|do(T=t),X,U,C) \\ & p(U|do(T=t),X)p(C|do(T=t),X)dUdC \\ \overset{*}{=} & \iint_{U,C} p(Y|T=t,X,U,C)p(U|X)p[C|X)dUdC. \end{aligned} \tag{3}$$

□

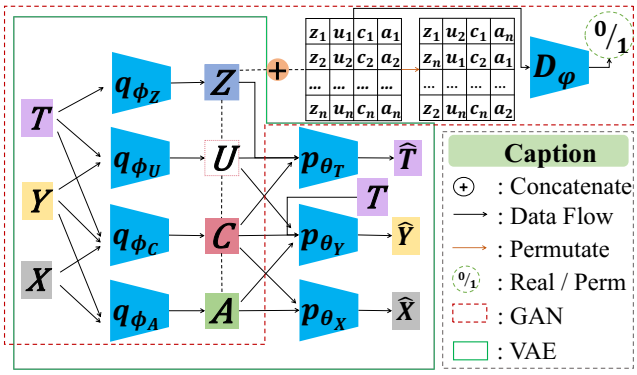

Figure 2: Overview of VIV Framework. During the max stage, we train the D_ψ to distinguish the real and permuted samples while fixing the parameters of other neural networks, e.q. q_ϕ and p_θ. During the min stage, we update the parameters of the inference and generative networks q_ϕ and p_θ of VAE while freezing the parameters of discriminator D_ψ. After min-max training, we save the latent factors of Z for downstream IV-based counterfactual prediction.

The equality * holds by the rules of *do-calculus* (Pearl et al. 2000). Next, we will show how we estimate $p(X, T, Y, U, Z, C, A)$ from the observations of (X, T, Y).

Bayesian Network of VIV

The causal graph of the education example (Hoxby 2000) is mapped onto a Bayesian network visually represented in Figure 1. Four exogenous variables in the original causal graph are translated into four latent factors in the Bayesian network. Furthermore, we construct a VIV framework corresponding to this Bayesian network, as depicted in Figure 2. This framework consists of an inference network q_ϕ, a generative network p_θ and a discriminator $\mathcal{D}_\psi$. Overall, the objective function of VIV Eq. (4) contains min-max stages for generating IVs and ensuring the latent factor disentanglement.

$$\begin{aligned} &\min_{\theta,\phi} \mathcal{L}_{VIV} = -ELBO_{VIV} + \alpha \mathcal{O}_{GAN}; \\ &\max_{\psi} \mathcal{O}_{GAN}. \end{aligned} \tag{4}$$

where the meanings and details of $ELBO_{VIV}$, TC and $\mathcal{O}_{\mathcal{D}_\psi}$ will be elaborated in the following sections.

Evidence Lower Bound of VIV

We derive the ELBO to uncover the latent factors in accordance with the Bayesian network structure presented in Figure 1. According to the local Markov assumption (Pearl et al. 2000), the joint distribution, $p_\theta(X, T, Y, U, Z, C, A)$, can be factorized as follows:

$$\begin{aligned} p_\theta(X, T, Y, U, Z, C, A) =& p(U) \times p(Z) \times p(C) \times p(A) \\ & p_\theta(X|C, A) \times p_\theta(T|Z, C, U) \\ & \times p_\theta(Y|A, C, U, T). \end{aligned} \tag{5}$$

Furthermore, we posit that the posterior distribution $q_\phi(U, Z, C, A|X, T, Y)$, can be factorized as follows:

$$\begin{aligned} q_\phi(U, Z, C, A|X, T, Y) = & q_\phi(Z|T) \times q_\phi(U|T, Y) \\ & \times q_\phi(A|X, Y) \times q_\phi(C|X, T, Y). \end{aligned} \tag{6}$$

Given (6), we obtain the variational lower bound as follows:

$$\begin{aligned} & log p_\theta(X, T, Y) \\ & \geq \mathbb{E}_{q_\phi(C|X,T,Y)q_\phi(A|X,Y)}[log p_\theta(X|C, A)] \\ & + \mathbb{E}_{q_\phi(Z|T)q_\phi(U|T,Y)q_\phi(C|X,T,Y)}[log p_\theta(T|Z, C, U)] \\ & + \mathbb{E}_{p(T)q_\phi(A|X,Y)q_\phi(U|T,Y)q_\phi(C|X,T,Y)} \\ & [log p_\theta(Y|A, C, U, T)] + KL(q_\phi(Z|T)||p(Z)) \\ & + KL(q_\phi(U|T, Y)||p(U)) + KL(q_\phi(A|X, Y)||p(A)) \\ & + KL(q_\phi(C|X, T, Y)||p(C)) \\ & \equiv ELBO_{VIV}, \end{aligned} \tag{7}$$

where KL denotes KL divergence (Kingma and Welling 2013). Detailed proof of (7) is provided in the supplementary materials. Based on the (7), the structures of encoder q_ϕ and decoder p_θ of VIV are designed as presented in Figure 2. In practice, we assume the prior and posterior distribution of latent factors follow the Gaussian distribution and adopt the reparametrization trick (Kingma and Welling 2013) for efficient gradient computation. Formally:

$$\begin{aligned} p(Z) &= \mathcal{N}(Z|0, I); q_\phi(Z|T) = \mathcal{N}(Z|\overline{\mu}_Z, \overline{\sigma}^2_Z); \\ p(U) &= \mathcal{N}(U|0, I); q_\phi(U|T, Y) = \mathcal{N}(U|\overline{\mu}_U, \overline{\sigma}^2_U); \\ p(A) &= \mathcal{N}(A|0, I); q_\phi(A|X, Y) = \mathcal{N}(A|\overline{\mu}_A, \overline{\sigma}^2_A); \\ p(C) &= \mathcal{N}(C|0, I); q_\phi(C|X, T, Y) = \mathcal{N}(C|\overline{\mu}_C, \overline{\sigma}^2_C); \\ \overline{\mu}_Z &= g^\mu_Z(T); \overline{\sigma}_Z = g^\sigma_Z(T); \\ \overline{\mu}_U &= g^\mu_U(T, Y); \overline{\sigma}_U = g^\sigma_U(T, Y); \\ \overline{\mu}_A &= g^\mu_A(X, Y); \overline{\sigma}_A = g^\sigma_A(X, Y); \\ \overline{\mu}_C &= g^\mu_C(X, T, Y); \overline{\sigma}_C = g^\sigma_C(X, T, Y); \end{aligned} \tag{8}$$

where I indicates the identity matrix.

For the generative network, we assume a Multinomial/Gaussian distribution for continuous/discrete variables. Here, we take the distributions of continuous variables as an example, Formally:

$$\begin{aligned} p_\theta(X|C, A) &= \mathcal{N}(X|\hat{\mu}_X, \hat{\sigma}^2_X); \\ p_\theta(T|Z, C, U) &= \mathcal{N}(T|\hat{\mu}_T, \hat{\sigma}^2_T); \\ p_\theta(Y|A, C, U, T) &= \mathcal{N}(Y|\hat{\mu}_Y, \hat{\sigma}^2_Y); \\ \hat{\mu}_X &= f^\mu_X(C, A); \hat{\sigma}_X = f^\sigma_X(C, A); \\ \hat{\mu}_T &= f^\mu_T(Z, C, U); \hat{\sigma}_T = f^\sigma_T(Z, C, U); \\ \hat{\mu}_Y &= f^\mu_Y(A, C, U, T); \hat{\sigma}_Y = f^\sigma_Y(A, C, U, T). \end{aligned} \tag{9}$$

Adversarial Learning of VIV

To realize the **Exogeneity** assumption of generated Z, we encourage the mutual independence between the latent factors, formally:

$$q(Z, U, C, A) = q(Z)q(U)q(C)q(A). \tag{10}$$

Intuitively, we are supposed to minimize the KL divergence $KL(q(Z,U,C,A)||q(Z)q(U)q(C)q(A))$, which is also known as *Total Correlation* (Watanabe 1960), a widely-used indicator of correlation among multiple random variables. However, the posterior distributions of latent factors are intractable as they are conditioned on observed variables. Inspired by (Kim and Mnih 2018), we adopt the *permutation trick* to approximate this KL divergence. For the sake of simplicity, we use q and $\overline{q}$ to denote $q(Z,U,C,A)$ and $q(Z)q(U)q(C)q(A)$, respectively. Specifically, we sample from $q(Z,U,C,A|X,T,Y)$ to get the real samples as a view for q, then randomly permuting across batches for each latent factor to obtain the permuted samples. Provided that the batch size is sufficiently large, the distribution of these samples will closely approximate the distribution $\overline{q}$ (Arcones and Gine 1992).

With the availability of samples from q and $\overline{q}$, we are able to use the density-ratio trick (Nguyen, Wainwright, and Jordan 2010) to narrow the gap between the two distributions. To elaborate further, we train a discriminator D_ψ to output the probability of the sample coming from $\overline{q}$ instead of q. In the max-stage, we train D_ψ to be discriminative, while in the min-stage, we fix the parameters of D_ψ and train the q_ϕ to generate latent factors with the distributions close to $\overline{q}$. Formally, the min-max objective $\mathcal{O}_{GAN}$ is as follows:

$$\begin{aligned}\min_{q_\phi}\max_{D_\psi}&\mathbb{E}_{Z,U,C,A\sim\overline{q}}[log(D_\psi(Z,U,C,A))]\\&+\mathbb{E}_{Z,U,C,A\sim q}[log(1-D_\psi(Z,U,C,A))].\end{aligned}\tag{11}$$

In the min-stage, the inference network q_ϕ is indeed playing a role of a **generative network for the latent factors from the marginal distribution** [2]. This is the main difference between our VAE-GAN-based framework with the classical VAE-GAN (Gur, Benaim, and Wolf 2020), where the whole VAE model is treated as the generative model for GAN. Besides, having low TC is only meaningful when we can retain information of the latent factors (Kim and Mnih 2018), we incorporate the Eq. (7) with TC loss to ensure the valid disentanglement.

Remark 0.5. *Reflecting on our model's design, we conclude that the following aspects guarantee the validity of the generated instrumental variables: (1) in the generative network p_θ, T is reconstructed directly from Z, guaranteeing the **Relevance**; (2) in the inference network q_ϕ, Y is involved in the inference of all the latent factors except Z. Indeed, the latent factors are categorized into two groups: directly related to outcomes, e.g., U,C,A, and indirectly related to outcomes, e.g., Z. The **Exclusivity** is thus achieved; (3) by diminishing the total relation TC between latent factors with an adversarial game, the correlation between Z and other factors is compelled to weaken, ensuring the **Exogeneity** of the generated instruments.*

In practice, we adopt gradient clipping trick (Zhang et al. 2019) and replace KL divergence with Wasserstein distance (Gulrajani et al. 2017) for stable training. Due to the page limitation, we leave the pseudo codes in the supplementary materials.

[2] q_ϕ is a generative network for latent factors from the marginal distribution and p_θ is a generative network for observed variables.

Experiments

Datasets

Simulated Dataset The **Demands** dataset is first constructed by (Hartford et al. 2017), which simulates the causal effects of price variation T on airline demands Y. In this example, the customer's type A represents different levels of price sensitivity. The holiday effect on sales is reflected with observed confounders, the time of the year C with a complex non-linear function κ_C. The conference hosting plays the role of unobserved confounders, the effects of which on demands are manifested by latent errors U. The parameter ρ is introduced to alter the correlation between T and U. A larger value of ρ signifies that the causal effects of T on Y will be more distorted by U. Fuel price is a typical **continuous** instrument Z in this example as it influences the demands only by ticket price.

Formally, the DGP is as follows:

$$\begin{aligned}Y&=100+(10+T)A\kappa_C-2T+U;\\T&=25+(Z+3)\kappa_C+V;\\\kappa_C&=2((C-5)^4/600+exp[-4(C-5)^2]+C/10-2);\\A&\in\{1,...,7\},C\in Unif(0,10);\\Z,&V\sim N(0,1),U\sim N(\rho V,1-\rho^2).\end{aligned}\tag{12}$$

In our settings, we can only access the IVs Candidates $X=[C,A]$, where Z is not obtainable due to some reasons, and we are not aware of the specific type of variables in X. In other words, we aim to recover the latent factors Z,U,C,A from the observed X,T,Y. We generate 10000/10000/10000 samples in the training/validation/testing dataset and perform 10 trials for each methods.

Real-World Dataset Following previous methods (Louizos et al. 2017), we conduct experiments on one real-world dataset: **Twins**. We pick up Z,C,A from the covariates in the dataset and generate U, T, and Y as presented in Eq. (12). To show the generality of our methods, we set Z as **discrete variables**. Following (Li and Yao 2022), the Twins dataset is divided 56/24/20 into training/validation/testing sets. We perform experiments for each baselines 10 times as well.

Baselines and Metrics

We compare our VIV with the state-of-the-art IV generation methods on the IV regression backbones.

IV Generation Methods We compare our VIV with the following IV generation methods: 1) NoneIV, which takes zero vectors as IVs; 2) Weighting methods: UAS, WAS (Burgess, Dudbridge, and Thompson 2016) and ModeIV (Hartford et al. 2021), which weight the IV candidates to generate IV synthesis; 3) Data-driven methods: AutoIV (Yuan et al. 2022) and GIV (Wu et al. 2023), which learn IV representations or group variables for IVs candidates; 4) TrueIV, which takes the true IVs in DGP for counterfactual prediction and is supposed to have the best performance.

In-Sample	Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
NoneIV	>1000	1.27(0.18)	0.52(0.04)	Nan	0.22(0.02)	0.31(0.05)	0.64(0.14)	1.27(0.05)	1.25(0.09)
UAS	>1000	1.20(0.14)	0.53(0.06)	6.19(2.02)	0.20(0.02)	0.29(0.05)	0.60(0.16)	1.26(0.04)	0.91(0.20)
WAS	>1000	1.13(0.13)	0.53(0.07)	5.99(3.38)	0.18(0.02)	0.32(0.03)	0.56(0.12)	1.28(0.05)	1.19(0.14)
ModeIV	>1000	1.20(0.14)	0.53(0.06)	6.19(2.02)	0.18(0.02)	0.29(0.05)	0.60(0.16)	1.26(0.04)	0.91(0.20)
AutoIV	6.14(5.40)	1.13(0.16)	0.53(0.05)	**2.97(1.01)**	0.20(0.04)	**0.26(0.05)**	0.62(0.18)	1.23(0.05)	1.10(0.27)
GIV-EM	2.89(1.92)	1.10(0.27)	0.54(0.09)	9.15(4.46)	0.23(0.03)	0.40(0.06)	0.35(0.21)	1.01(0.07)	0.99(0.09)
VIV	**0.45(0.05)**	**0.39(0.10)**	**0.36(0.06)**	**2.44(0.62)**	**0.17(0.01)**	0.34(0.03)	**0.18(0.02)**	**0.53(0.05)**	**0.40(0.05)**
TrueIV	**0.18(0.02)**	**0.13(0.03)**	**0.20(0.03)**	4.35(1.35)	**0.10(0.02)**	**0.17(0.03)**	**0.09(0.02)**	**0.24(0.01)**	**0.16(0.02)**
Out-Sample	Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
NoneIV	>1000	1.26(0.17)	0.52(0.04)	Nan	0.22(0.02)	0.31(0.05)	0.63(0.15)	1.26(0.04)	1.25(0.09)
UAS	>1000	1.20(0.14)	0.52(0.05)	6.19(1.98)	0.20(0.01)	0.29(0.04)	0.60(0.17)	1.26(0.04)	0.91(0.20)
WAS	>1000	1.12(0.11)	0.53(0.06)	6.06(3.38)	0.18(0.02)	0.32(0.03)	0.55(0.12)	1.27(0.04)	1.18(0.14)
ModeIV	>1000	1.20(0.14)	0.52(0.05)	6.19(1.98)	0.18(0.02)	0.29(0.04)	0.60(0.17)	1.26(0.04)	0.91(0.19)
AutoIV	6.19(5.49)	1.12(0.16)	0.53(0.04)	**3.00(1.02)**	0.19(0.03)	**0.26(0.05)**	0.61(0.18)	1.23(0.04)	1.10(0.26)
GIV-EM	2.86(1.90)	1.10(0.28)	0.53(0.09)	9.26(4.50)	0.23(0.04)	0.40(0.06)	0.35(0.21)	1.01(0.08)	0.98(0.10)
VIV	**0.45(0.05)**	**0.39(0.10)**	**0.36(0.06)**	**2.46(0.62)**	**0.17(0.01)**	0.34(0.03)	**0.17(0.02)**	**0.52(0.05)**	**0.40(0.05)**
TrueIV	**0.18(0.02)**	**0.13(0.03)**	**0.20(0.03)**	4.35(1.35)	**0.10(0.02)**	**0.17(0.03)**	**0.09(0.02)**	**0.24(0.01)**	**0.16(0.02)**

Table 1: Performance comparison of MSE of the counterfactual prediction on do(T) outcomes between VIV and the SOTA baselines on the Demands-0.5 datasets. Bold indicates the method with the best and second-best performance.

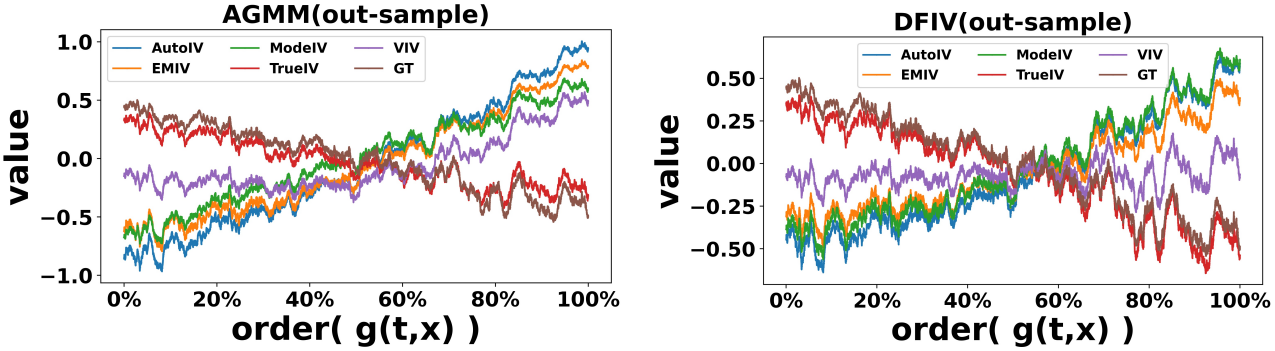

Figure 3: Counterfactual prediction curves with different IV-based generation methods. Compared to other IV-generation methods, the counterfactual prediction curve based on VIV closely approximates the ground truth curve and the counterfactual prediction curve based on TrueIV.

IV Regression Backbones We plug the generated IVs into the downstream IV-based counterfactual prediction algorithms to check the validity of the generated IVs. These backbones include: 1) 2SLS-based methods: Poly2SLS, NN2SLS (Angrist and Imbens 1995); 2) Kernel-based methods: KernelIV (Singh, Sahani, and Gretton 2019), DualIV (Muandet et al. 2020); 3) Deep methods: DeepIV (Hartford et al. 2017), OneSIV (Lin et al. 2019), DFIV (Xu et al. 2020) 4) GMM-based methods: DeepGMM (Bennett, Kallus, and Schnabel 2019), AGMM (Dikkala et al. 2020).

Metrics To answer the counterfactual question, "What would Y have been if T has been changed to t?", we create interventions by setting the value of T deterministically across a predefined grid of values covering the entire range of T in the training set when evaluating the performance of the methods. We take Mean Squared Error (MSE) as the evaluation metric to compare the predicted outcomes after intervention and ground truth.

Results

Results on Demands Table 1 shows the performance of VIV on a simulated Demands-0.5 dataset, where 0.5 denotes the value of ρ in Eq. (12). We compare our method with the 6 IV-generation methods on 9 backbone IV-based counterfactual prediction models. We also test the performance of true instrumentals $TrueIV$ on backbones as the baseline for

In-Sample	Poly2SLS	NN2SLS	KernelIV	DeepIV	OneSIV	DeepGMM	AGMM	Average
NoneIV	10.3(27.5)	0.50(0.11)	0.36(0.10)	**0.34(0.04)**	0.24(0.02)	0.39(0.08)	0.55(0.09)	1.81(3.74)
UAS	1.29(0.90)	0.52(0.11)	0.35(0.10)	0.34(0.05)	0.24(0.03)	0.43(0.07)	0.49(0.07)	0.52(0.35)
WAS	1.52(0.92)	0.41(0.07)	0.33(0.11)	0.35(0.04)	0.26(0.02)	0.37(0.10)	0.46(0.06)	0.53(0.44)
ModeIV	1.29(0.90)	0.52(0.11)	0.35(0.10)	0.35(0.03)	0.24(0.03)	0.43(0.07)	0.49(0.07)	0.52(0.35)
AutoIV	19.8(37.8)	0.49(0.13)	**0.32(0.09)**	0.37(0.02)	0.29(0.04)	0.41(0.10)	0.57(0.13)	3.18(7.33)
GIV-EM	9.30(20.9)	0.38(0.07)	0.35(0.11)	0.35(0.03)	0.29(0.03)	0.48(0.16)	0.66(0.08)	1.69(3.36)
VIV	**0.17(0.05)**	**0.23(0.06)**	0.36(0.25)	0.34(0.05)	**0.17(0.02)**	**0.34(0.07)**	**0.31(0.09)**	**0.27(0.08)**
TrueIV	**0.21(0.10)**	**0.16(0.04)**	**0.17(0.08)**	**0.23(0.05)**	**0.10(0.02)**	**0.21(0.02)**	**0.18(0.03)**	**0.18(0.04)**

Table 2: Performance comparison of MSE of the counterfactual prediction on do(T) outcomes between VIV and the SOTA baselines on the Twins datasets. Bold indicates the method with the best and second-best performance.

comparison. The mean and standard deviation of the MSE are reported outside and inside the parentheses, respectively.

We can have the following findings from the results in Table 1: 1) Weighting methods, e.g., UAS, WAS and ModeIV, lack reliability and cannot generate a valid IV in the absence of valid IV candidates. Incorporating them into IV methods scarcely enhances estimation performance, resulting in outcomes similar to those achieved without IVs (NoneIV). 2) For data-driven methods, the performance of AutoIV has suffered significantly as it can no longer learn effective IVs representations from features without valid IVs. The poor performance of GIV-EM demonstrates its limited application, as it emphasizes learning group IVs from diverse treatment assignment mechanisms. 3) VIV and AutoIV perform better than TrueIV with the DualIV method. This could be attributed to the fact that the DualIV approach necessitates continuous IVs, a condition met by the generated representations of AutoIV and latent factors of VIV. 4) Overall, when true instruments are **continuous variables**, VIV achieves the closest performance with $TrueIV$ on most backbone methods compared with other IV-generation methods in both in-sample and out-sample settings, manifesting the effectiveness of our method on generating valid IVs in the absence of valid IVs candidates.

Counterfactual Prediction Visualization To visualize the validness of our generated IVs, we plot the graphs depicting the estimated values of the effect function using the intervention T=do(t) on two SOTA IV-based counterfactual prediction backbones and arrange it based on the actual observed outcomes (Ground Truth, GT) on Demands-0.5 dataset. As shown in Figure 3, compared with other IV-generation methods, our VIV exhibits a trend that closely aligns with the curves of both the ground truth values and the prediction values based on real instrumental variables. Despite the remaining gap from the GT and TrueIV curves, there is a noticeable improvement in comparison to other methods with this new IV-generation approach, which indicates that we are making substantial progress in the right direction for advancing counterfactual prediction.

Results on Twins To show the **generality** of our IVs generation method, we conduct extensive experiments on a real-world dataset Twins, where the True IV are set as **discrete variables**. We present the related experimental results on in-sample setting in Table 2. We compute each IVs generation method's mean value and standard deviation of MSE on different backbones and report them in the **Average** column.

The results on the Twins dataset yield the following insights: 1): AutoIV and GIV perform worst among all baselines. This might be attributed to the limited sample size of the Twins dataset, which restricts the data-driven method from fully capturing the underlying causal relationships. Our framework exhibits a more **efficient** capacity for extracting causal features compared to other data-driven methods. 2): VIV possesses the lowest mean and variance in average counterfactual prediction performance, which demonstrates the most **stable** capability for valid instruments generation of our method. 3): VIV achieves excellent performance under both discrete and continuous settings of real instrumental variables. This achievement surpasses any comparative methods, showcasing the **generality** of the instrumental variables generated by our approach.

Due to the page limitation, we leave the hardware environment used for the experiment and optimal hyper-parameters in supplementary materials. The project page with the code and the supplementary materials is available at **GitHub**[3].

Conclusion

To resolve the challenge of generating instrumental variables in the absence of valid instrument candidates, we construct an adversarial variational autoencoder to generate valid instruments in the latent space. Extensive experimental results on synthetic and real-world datasets validate the effectiveness, stability and generality of our proposed framework compared the cutting-edged IV generation methods.

Due to the lack of real-world counterfactual datasets, VIV is only evaluated on limited tasks. It would be interesting to extend our method into other research areas, such as large language models (Kıcıman et al. 2023) and reinforcement learning (Ding et al. 2022) for robust feature extraction. We leave future work for the generalization of our method to these fields.

[3]https://github.com/XinshuLI2022/VIV

References

Alaa, A. M.; and van der Schaar, M. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30.

Angrist, J.; and Imbens, G. 1995. Identification and estimation of local average treatment effects.

Arcones, M. A.; and Gine, E. 1992. On the bootstrap of U and V statistics. *The Annals of Statistics*, 655–674.

Bennett, A.; Kallus, N.; and Schnabel, T. 2019. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32.

Burgess, S.; Dudbridge, F.; and Thompson, S. G. 2016. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11): 1880–1906.

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. 2013. Inference on counterfactual distributions. *Econometrica*, 81(6): 2205–2268.

Davies, N. M.; von Hinke Kessler Scholder, S.; Farbmacher, H.; Burgess, S.; Windmeijer, F.; and Smith, G. D. 2015. The many weak instruments problem and Mendelian randomization. *Statistics in medicine*, 34(3): 454–468.

Dikkala, N.; Lewis, G.; Mackey, L.; and Syrgkanis, V. 2020. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33: 12248–12262.

Ding, W.; Lin, H.; Li, B.; and Zhao, D. 2022. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. *Advances in Neural Information Processing Systems*, 35: 26532–26548.

Dupont, E. 2018. Learning Disentangled Joint Continuous and Discrete Representations. In *Neural Information Processing Systems*.

Fortuin, V.; Baranchuk, D.; Rätsch, G.; and Mandt, S. 2020. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, 1651–1661. PMLR.

Glass, T. A.; Goodman, S. N.; Hernán, M. A.; and Samet, J. M. 2013. Causal inference in public health. *Annual review of public health*, 34: 61.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Gur, S.; Benaim, S.; and Wolf, L. 2020. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33: 16761–16772.

Hartford, J.; Lewis, G.; Leyton-Brown, K.; and Taddy, M. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, 1414–1423. PMLR.

Hartford, J. S.; Veitch, V.; Sridhar, D.; and Leyton-Brown, K. 2021. Valid causal inference with (some) invalid instruments. In *International Conference on Machine Learning*, 4096–4106. PMLR.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Hoxby, C. M. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review*, 90(5): 1209–1238.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *International Conference on Machine Learning*.

Kim, H.; Shin, S.; Jang, J.; Song, K.; Joo, W.; Kang, W.; and Moon, I.-C. 2021. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8128–8136.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114.

Kuang, Z.; Sala, F.; Sohoni, N.; Wu, S.; Córdova-Palomera, A.; Dunnmon, J.; Priest, J.; and Ré, C. 2020. Ivy: Instrumental variable synthesis for causal inference. In *International Conference on Artificial Intelligence and Statistics*, 398–410. PMLR.

Li, S.; Vlassis, N.; Kawale, J.; and Fu, Y. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *IJCAI*, 3768–3774.

Li, X.; and Yao, L. 2022. Contrastive Individual Treatment Effects Estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*, 1053–1058.

Lin, A.; Lu, J.; Xuan, J.; Zhu, F.; and Zhang, G. 2019. One-stage deep instrumental variable method for causal inference from observational data. In *2019 IEEE International Conference on Data Mining (ICDM)*, 419–428. IEEE.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Muandet, K.; Mehrjou, A.; Lee, S. K.; and Raj, A. 2020. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33: 2710–2721.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19: 2.

Pfohl, S. R.; Duan, T.; Ding, D. Y.; and Shah, N. H. 2019. Counterfactual reasoning for fair clinical risk prediction.

In *Machine Learning for Healthcare Conference*, 325–358. PMLR.

Puli, A.; and Ranganath, R. 2020. General Control Functions for Causal Effect Estimation from IVs. *Advances in neural information processing systems*, 33: 8440–8451.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Singh, R.; Sahani, M.; and Gretton, A. 2019. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1): 66–82.

Wooldridge, J. M. 2015. Control Function Methods in Applied Econometrics. *Journal of Human Resources*, 50(2): 420–445.

Wu, A.; Kuang, K.; Li, B.; and Wu, F. 2022. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, 24056–24075. PMLR.

Wu, A.; Kuang, K.; Xiong, R.; Zhu, M.; Liu, Y.; Li, B.; Liu, F.; Wang, Z.; and Wu, F. 2023. Learning Instrumental Variable from Data Fusion for Treatment Effect Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10324–10332.

Wu, P. A.; and Fukumizu, K. 2022. β-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap. In *International Conference on Learning Representations*.

Xu, L.; Chen, Y.; Srinivasan, S.; de Freitas, N.; Doucet, A.; and Gretton, A. 2020. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Yuan, J.; Wu, A.; Kuang, K.; Li, B.; Wu, R.; Wu, F.; and Lin, L. 2022. Auto iv: Counterfactual prediction via automatic instrumental variable decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4): 1–20.

Zhang, J.; He, T.; Sra, S.; and Jadbabaie, A. 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.

Zimmermann, R. S.; Sharma, Y.; Schneider, S.; Bethge, M.; and Brendel, W. 2021. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, 12979–12990. PMLR.