Specify Robust Causal Representation from Mixed Observations

Mengyue Yang
University College London
London, United Kingdom
mengyue.yang.20@ucl.ac.uk

Xinyu Cai
Nanyang Technological University
Singapore, Singapore
xinyu.cai@ntu.edu.sg

Furui Liu *
Zhejiang Lab
Hangzhou, China
liufurui@zhejianglab.com

Weinan Zhang
Shanghai Jiao Tong University
Shanghai, China
wnzhang@apex.sjtu.edu.cn

Jun Wang
University College London
London, United Kingdom
j.wang@cs.ucl.ac.uk

ABSTRACT

Learning representations purely from observations concerns the problem of learning a low-dimensional, compact representation which is beneficial to prediction models. Under the hypothesis that the intrinsic latent factors follow some casual generative models, we argue that by learning a causal representation, which is the minimal sufficient causes of the whole system, we can improve the robustness and generalization performance of machine learning models. In this paper, we develop a learning method to learn such representation from observational data by regularizing the learning procedure with mutual information measures, according to the hypothetical factored causal graph. We theoretically and empirically show that the models trained with the learned causal representations are more robust under adversarial attacks and distribution shifts compared with baselines. The supplementary materials are available at https://github.com/ymy4323460/CaRI/.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

Causal representation learning, Robustness learning, Learning theory for Causal representation learning.

ACM Reference Format:
Mengyue Yang, Xinyu Cai, Furui Liu, Weinan Zhang, and Jun Wang . 2023. Specify Robust Causal Representation from Mixed Observations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3580305.3599512

1 INTRODUCTION

Causal representation learning is an effective approach for extracting invariant, cross-domain stable causal information, which is believed to be able to improve sample efficiency by understanding the underlying generative mechanism from observational data [3, 30]. Causal representation learning is widely applied in many real-world applications like recommendation systems, search engines etc.[22, 35, 39, 46]. Recently, multiple approaches were proposed to learn the invariant causal representations, which are supposed to encode underlying causal generative systems describing the data, based on the problem-specific priors.

The usual theory to implement it is called Independent Causal Machine (ICM) [24] principle, which can be applied to identify the cause information when all factors are observable. However, when the variables are unobservable in general and complex systems, this method usually does not work. Given that most methods employ a generative model, the main reason for such failure is due to the observation data (e.g. human images) is entangled by causal variables. To tackle this problem, previous works learned latent representations to capture the causal properties, e.g., causal disentanglement methods [33, 44] and invariant causal representation learning method [2, 17]. However, additional information like causal variable labels and domain information should be provided, which is usually unavailable in real-world systems.

In this paper, we aim at disentangling the causal variables from an information theoretical view without providing additional supervision signals. Supposing that the factors are casually structured, we formalize a causal system as in Fig.1 (a), which is commonly accepted by the causality community [40, 42]. Given the label Y, the d-dimensional observational data $\mathbf{X}$ is consist of causal factors including the parents $\mathbf{pa_Y}$, non-descendants $\mathbf{nd_Y}$, descendants $\mathbf{dc_Y}$ of $\mathbf{Y}$. The causal information $\mathbf{pa_Y}$ enables the model a better generalization and robustness for prediction tasks. We consider the natural data generative process as an information propagation along the causal graph and try to find out $\mathbf{pa_Y}$ from $\mathbf{X}$. Based on the causal modelling, we propose to learn latent representations which maintain the most necessary causal information for the prediction task, named minimal sufficient causal information of a system.

More specifically, we define the minimal sufficient cause (MSC) $\mathbf{Z}$ as a proxy of the parents in factor space as shown in Fig. 1 (b). MSCs are variables that are specially positioned in the system, blocking the path from the causes and non-descendants to Y. In this paper, we implement it by an information-theoretical approach, reducing the traditional two-step procedure i.e. causal disentanglement and information minimizing, to an optimization problem that can directly learn a latent causal representation with minimal sufficiency from observations. Specifically, the proposed optimization problem is a bi-level optimization problem minimizing $I(\mathbf{Z}; \mathbf{pa_Y}, \mathbf{nd_Y})$, with

[*] Corresponding Author

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599512

maximizing mutual information $I(\mathbf{Z}; Y)$ as a constraint. Based on this, we propose an intervention effect to accurately specify the causal information $\mathbf{pa_Y}$. We name this method as **CaRI** (learning Cause Representation by Information-theoretic approach) and we further extend the method under robustness learning framework. Moreover, we theoretically analyze the sample efficiency of CaRI by giving a generalization error bound with respect to sample size. Experiments on synthetic and real-world datasets show the effectiveness of the proposed method.

The main contribution of this paper are summarized below:

- We define minimal sufficient causes (MSC) in causal system by the formalization of an explicit causal graphical model to describe the data generative process of the real-world system and propose an information-theoretical approach to learn MSC from observational data.
- We theoretically analyze the sample efficiency of the learning approach by giving a generalization error bound w.r.t sample size. The theorem depicts a quantitative link between the amount of causal information contained in the learned representation and the sample complexity of the model on downstream tasks.
- We empirically verify that CaRI is able to generalize well distribution shift respectively and robust against adversarial attack.

2 RELATED WORKS

Causal Representation Learning is a set of approaches to finding generalizable representations by extracting and utilizing causal information from observational data. They usually aim at finding causal structure and causal variables behind observations. From several different perspectives, a bunch of methods have been proposed in the literature.

Causal Structure Learning. To assess the connection between causally related variables in real-world systems, a bunch of traditional methods use the Markov condition and conditional independence between cause and mechanism principle (ICM) to discover the causal structure or distinguish causes from effect [21]. Several works focus on the asymmetry between cause and effect. [9, 12, 37, 38], and similar ideas are utilized by [24, 38]. The series of works always assume that all the variable is observable. In contrast with these works, our proposed method is applicable to scenarios where the observed data is generated by hidden causal factors.

Invariant Representation Learning Cross Multidomain. Some pioneering work [34, 42, 47] considers the heterogeneity across multiple domains under the out-of-distribution settings [10, 16, 17, 19, 20, 26, 29, 45]. They learn causal representations from observational data by enforcing invariant causal mechanisms between the causal representation and the task labels across multi-domains. Similar to these works, we target obtaining invariant latent causal information but do not assume that the datasets are collected from multi-domains.

Causal Disentanglement Representation Learning. Causal representation learning helps to reduce the dimension of the original high-dimensional input. Several works leverage structural causal models to describe causal relationships inside the entangled observational data [33, 42, 44] and learn to disentangle causal concepts from original inputs. Different from aforementioned works, the proposed method in this paper considers the causal information from the perspective of information theory [4, 7]. We put our attention on minimal causal information, which can be regarded as a compact representation of the whole underlying causal system. We also theoretically analyze the generalization ability from PAC learning frameworks [31, 32] and explain why the causal representation can achieve better generalization ability from the perspective of sample complexity.

3 PROBLEM DEFINITION

3.1 Notations

Considering the causal scenario in Fig.1 (a), the observation data can be generated by the concepts in hidden space which contain multiple hidden causal variables. Denote $\mathbf{X} \in \mathcal{X}$ as d-dimensional observational data like context information or features in real-world systems, and $Y \in \mathcal{Y}$ as the labels of downstream tasks. Each pair of sample $(\mathbf{x},y)$ is drawn i.i.d. from joint distribution $p(\mathbf{x}, y)$. We use $\mathbf{pa_Y} \in \mathbb{R}^{p_1}$ to denote the variables including parent nodes of Y in the causal graph, while $\boldsymbol{\epsilon}$ is the vector of independent noise with probability densities of $p_{\boldsymbol{\epsilon}} = \mathcal{N}(0, \beta I)$. Similarly, $\mathbf{dc_Y} \in \mathbb{R}^{p_2}$ and $\mathbf{nd_Y} \in \mathbb{R}^{p_3}$ denote the descendant and non-descendant nodes of $\mathbf{Y}$, respectively. In our method, we introduce minimal sufficient parents, denoted by $\mathbf{Z} \in \mathcal{Z}$ of the system. Note that all the causal factors are assumed to be embedded in factors space, the observed data only contains $(\mathbf{X}, Y)$, where $\mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}), \mathbf{h} \in \mathcal{H}$ where $\mathbf{h} : \mathbb{R}^{p1+p2+p3} \to \mathbb{R}^d$ is a deterministic function. In causal systems, the causes of prediction tasks are stable and robust, this means that when intervening on the parents, the causal effect is propagated to its child but not vice versa. All other correlated variables $\mathbf{nd_Y}, \mathbf{dc_Y}$ in the causal system are regarded as spurious-correlated variables.

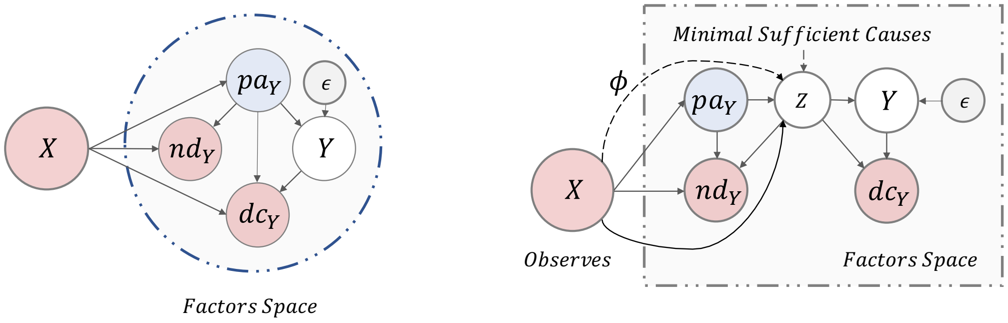
3.2 Minimal Sufficient Causes (MSC)

In our paper, we claim that not all the cause information is useful for prediction tasks. For example, considering a case of burning fire in a room, it is the presence of oxygen which explain the fire, but the match struck is definitely the necessary cause of fire. This real-world example is selected from section 9 in [25]. From the perspective of finding the most useful causes from observational data, we introduce the minimal sufficient cause variable $\mathbf{Z}$ into the causal system. As Fig. 1 (b) shows, the minimal sufficient causes $\mathbf{Z}$ are regarded as the proxy of parent variables. We define minimal sufficient causes in detail as below.

Definition 1. Assuming that the causal graph (Fig. 1 (b)) with Minimal Sufficient Causes holds, the Minimal Sufficient Cause blocks the path between $[\mathbf{pa_Y}, \mathbf{nd_Y}]$ and Y, and the following conditional independence condition holds:

$$(\mathbf{pa_Y}, \mathbf{nd_Y}) \perp Y | \mathbf{Z} \tag{1}$$

Our goal is to identify the minimal sufficient information $\mathbf{Z}$ in hidden factors space. The minimum sufficient causal variable $\mathbf{Z}$ in a causal system is stable information for predicting y. From the perspective of sufficient causes, we define it from a probabilistic view, which is inspired by the minimal sufficient statistics [15].



(a) *general causal system*

(b) *causal system with Minimal Sufficient Causes*

Figure 1: The figure demonstrates a case of a causal system (a) and its extension of introducing minimal sufficient causes (b).

Definition 2. (Sufficient Causes). Let $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{Z}, Y$ be random variables. $\mathbf{Z}$ is sufficient cause of Y shown in Fig. 1 (b) if and only if

$$\forall \mathbf{pa_y} \in \mathbf{pa_Y}, \mathbf{nd_y} \in \mathbf{nd_Y}, y \in Y; p(y|\mathbf{z}, \mathbf{pa_y}, \mathbf{nd_y}) = p(y|\mathbf{z}) \quad (2)$$

The definition of sufficient causes are the variables that are able to "produce" the causal system. From the perspective of minimal, we define a variable which can generate the whole the system with the minimum information. That is, all the variables of prediction task can be inferred if minimal sufficient causes are given.

Definition 3. (Minimal Sufficient Causes). The sufficient cause $\mathbf{Z}^*$ is minimal if and only if for any sufficient cause $\mathbf{Z}$, there exists a deterministic function $\mathbf{f}$ such that $\mathbf{Z}^* = \mathbf{f}(\mathbf{Z})$ almost everywhere w.r.t. $\mathbf{X}$.

Definition 3 shows that the Minimal Sufficient Causes $\mathbf{Z}$ in the causal system is the variable containing minimal information from all parents.

3.3 Learning MSC as Causal Representation from Observational Data

This paper focuses on causal representation learning, which aims at finding a low-dimensional representation of observation benefiting for predicting Y. Fig. 1 (a)(b) shows the causal system behind a prediction task, which uses observational data $\mathbf{X}$ to predict the target Y. The method is treated as a two stage process, and the first stage is to extract the representation from observational data. Let $\mathbf{Z} = \phi(\mathbf{X})$ denote representation extracted from original observation $\mathbf{X}$, where $\phi : \mathcal{X} \to \mathcal{Z}$ is the representation extraction function. The next stage is to use the representation to predict Y.

Now that we have formally defined Minimal Sufficiency, the basic objective is defined as learning a representation where all the information from minimal sufficient causes is included. The process is to model a flow of representation learning method and downstream prediction by satisfying Definition 2 3. The objective from Definition 2 is easy to be evaluated by common statistic methods, like independent testing by mutual information. However, it is very hard to get the minimal variable in Definition 3. To evaluate the objectives in Definition 2 3 in a unified framework, we utilize the information-theoretic ways since it can naturally combine Definition 2 and 3 by considering the information contained in MSC.

4 LEARNING MINIMAL SUFFICIENT CAUSAL REPRESENTATIONS

In this section, we present a method to learn the minimal sufficient parent's information $\mathbf{Z}$ from observational data $\mathbf{X}$. The difficulty lies in distinguishing minimal sufficient cause $\mathbf{Z}$ from $\mathbf{X}$, when we only observe $\mathbf{X}$. We first analyze the information propagation among different causal variables under two typical causal graphs in hidden factors space, based on which we propose an objective function with mutual information constraints. Next, we extend our method by introducing do-operation, which can enhance the ability to distinguish causes if such information is not embedded in the observational data.

4.1 Information-theoretic property of MSC in factor space

An important fact is that in Fig.1 (b), the minimal sufficient causes in observational data $\mathbf{X}$ dominate the generative process of the causal system defined in Fig.1 (b). If there exists a mapping from $\mathbf{X}$ to $\mathbf{Z}$, it is a function that finds the minimal sufficient causes inside the causal system. We develop an algorithm to learn representations based on such hypothetical structure Fig. 1 (b). Based on the definition of $\mathbf{Z}$, denoted by $I(\cdot, \cdot)$ the mutual information, we obtain the following Theorem (The proof is provided in supplementary material).

Theorem 4.1. *Let $\mathbf{Z} \in \mathcal{Z}$, $\mathbf{Z} = \phi(\mathbf{X})$, $\mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y})$ and $\mathbf{h} \in \mathcal{H}$ is an invertible function, $\mathbf{Z}$ is a minimal sufficient cause of the causal system demonstrated in Fig. 1 (b) if and only if $\mathbf{Z}$ is an optimal solution of following objective*

$$\begin{aligned} &\min_{\mathbf{Z}} I(\mathbf{Z}; \mathbf{pa_Y}, \mathbf{nd_Y}) \\ &s.t.\ \mathbf{Z} \in \arg\max_{\mathbf{Z}'} I(\mathbf{Z}'; Y) \end{aligned} \quad (3)$$

Theorem 4.1 shows that we can identify the MSC by solving the min-max optimization problem. In real-world applications, the information of $\mathbf{nd_Y}$ and $\mathbf{dc_Y}$ may not be revealed, and the above objective function cannot be optimized directly. To get a tractable form, in the next section, we extend our optimization objective to observational space. We extend Eq. 3 to a tractable objective by scaling the mutual information terms in Eq. 3. The way is to link the unrevealed variables $\mathbf{nd_Y}, \mathbf{dc_Y}$ to observation $\mathbf{X}$. The following lemma can help us scale Eq. 3.

Lemma 4.2. *Suppose the features and labels are* $\mathbf{X}, Y$ *respectively, where* $\mathbf{X}$ *deterministically consists of the minimal sufficient parents, descendants and non-descendant as* $\mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y})$. *The following inequality holds if and only if* $\mathbf{h}$ *is an invertible deterministic function*

$$I(\mathbf{Z}; \mathbf{nd_Y}, \mathbf{pa_Y}) \leq I(\mathbf{Z}; \mathbf{X}) \tag{4}$$

Proposition 4.3. *Let* $\mathbf{Z}', \mathbf{Z} \in \mathcal{Z}, \mathbf{Z} = \phi(\mathbf{X}), \mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y})$, $h \in \mathcal{H}$ $\mathbf{h}(\cdot)$ *is invertible function. When all the functions (lines in Fig. 1) between* $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$ *are invertible* $\mathbf{Z}$ *is the minimal sufficient cause of the causal system demonstrated in Fig. 1 (b) if and only if* $\mathbf{Z}$ *equals to the optimal solution of following objective*

$$\begin{aligned} &\min_{\mathbf{Z}} I(\mathbf{Z}; \mathbf{X}), \\ &s.t.\ \mathbf{Z} \in \arg\max\nolimits_{\mathbf{Z}'} I(\mathbf{Z}'; Y) \end{aligned} \tag{5}$$

From Theorem 4.3 we can substitute the terms including $\mathbf{nd_Y}$ and $\mathbf{pa_Y}$ by tractable mutual information term. For Eq. 5 in Proposition 4.3, by defining $m(Y) = \max_{\mathbf{Z}'} I(\mathbf{Z}'; Y)$ and then reformulating $\mathbf{Z} \in \arg\max_{\mathbf{Z}'} I(\mathbf{Z}'; Y)$ as $I(\mathbf{Z}; Y) \geq m(Y)$, with plugging in $\mathbf{Z} = \phi(\mathbf{X})$ the optimization problem in Eq. 3 can be equivalently formulated as minimizing the following Lagrangian $\mathcal{L}(\phi, \lambda)$ such that

$$\delta(\phi) = \mathcal{L}(\phi, \lambda) = I(\phi(\mathbf{X}); \mathbf{X}) - \lambda I(\phi(\mathbf{X}); Y), \tag{6}$$

where $\mathcal{L}(\phi, \lambda)$ is denoted as $\delta(\phi)$ for conciseness and λ is a hyperparameter which is manually selected in practice. The object coincides with Information Bottleneck (IB) objective function [32]. The difference is that IB is deduced from Rate Distortion Theorem in information theory, and it holds under the structure of the Markov Chain instead of a causal graph (i.e. Fig. 1). In this paper, the IB setting is generalized into causal space, by bridging minimal sufficient causes with root cause variables in the hypothetical causal graph. The detailed proof of Theorem 4.1 and Proposition 4.3 in supplementary materials show the differences between our proposed method and IB.

4.2 Distinguishing components by intervention effect

The previous section illustrates a method to find $\mathbf{Z}$ on the factor and observation level. Note that the objective function $\delta(\phi)$ (Eq. **??**) given from Proposition 4.3 can find the minimal sufficient causes under the strong assumption. In real-world applications, if we use the information-theoretic objective, it is very hard to distinguish causes $\mathbf{pa_Y}$ and spurious variable $\mathbf{dc_Y}, \mathbf{nd_y}$ of Y from the objective. To release this problem, we introduce an intervention operation, denoted by $do(X = x)$ [25] into our method. Intervention in causality means the system operates under the condition that certain variables are controlled by external forces. In hidden factor space, one of the differences between $\mathbf{Z}$ and $\mathbf{dc_Y}, \mathbf{nd_y}$ is that if we intervene on the value of $\mathbf{Z}$, the causal effect will be delivered to its child Y, but the causal effect to Y from the intervention conducted on child node $\mathbf{dc_Y}$ will be blocked. From such, let $\bar{\mathbf{x}}$ means the intervened value which not equals to $\mathbf{x}$, the following inequality describes intervention effect holds

$$\begin{aligned} &P(Y = y|do(\mathbf{Z} = \mathbf{z})) - P(Y = y|do(\mathbf{Z} = \bar{\mathbf{z}})) > \\ &P(Y = y|do(\mathbf{dc_Y} = \mathbf{dc_y}))) - P(Y = y|do(\mathbf{dc_Y} = \bar{\mathbf{dc_y}})) = \\ &P(Y = y|do(\mathbf{nd_Y} = \mathbf{nd_Y}))) - P(Y = y|do(\mathbf{nd_Y} = \bar{\mathbf{nd_Y}})) = 0 \end{aligned} \tag{7}$$

Instead of conducting interventions on the parental variables in a real-world environment, we create a representation space $\mathcal{Z}$ where it supports simulation of the interventional manipulation on parents by intervening $\mathbf{Z}$ in the learned model. The functional interventional distributions $P(Y = y|do(\mathbf{Z} = \phi(\hat{\mathbf{x}})))$ can be identified from purely observational data $\mathbf{X}$ and Y ([25, 27, 42]),

$$\begin{aligned} &P(Y = y|do(\mathbf{Z} = \phi(\hat{\mathbf{x}}))) \\ &= \int_{\mathbf{x}} P(Y = y|\mathbf{x}, \hat{\mathbf{z}})|_{\hat{\mathbf{z}}=\phi(\hat{\mathbf{x}})} P(\hat{\mathbf{z}}|\mathbf{pa_y}) P(\mathbf{pa_y}|\mathbf{x}) P(\mathbf{X} = \mathbf{x}) d\mathbf{x} \\ &= E_{\mathbf{x}}[P(Y = y|\hat{\mathbf{z}})]|_{\hat{\mathbf{z}}=\phi(\hat{\mathbf{x}})} \end{aligned} \tag{8}$$

Therefore in the representation space, we can directly maximize the intervention effect on the intervention space $\mathcal{Z}$ to satisfy Eq. 7. To make the intervention effect easier to be evaluated in the mutual information process, we introduce an intervention variable $\bar{\mathbf{Z}} \in |\mathcal{Z}|$ and build an intervention network shown in Fig. 2, in which we first infer the representation $\mathbf{z}$ from observational data $\mathbf{x}$, based on which we can obtain the intervened value $\bar{\mathbf{z}} \neq \mathbf{z}$. Then we optimize the parameters in the model by maximizing the intervention effect term defined in mutual information language by

$$\begin{aligned} \text{Intervention Effect} &= \int_{\mathbf{z}, y} p(\mathbf{z}, y) \log p(y|\mathbf{z}) dy d\mathbf{z} \\ &- \int_{y, \mathbf{z}} p(y, \bar{\mathbf{z}}) \log p(y|\bar{\mathbf{z}}) dy d\bar{\mathbf{z}} \\ &= I(\mathbf{Z}; Y) - I(\bar{\mathbf{Z}}; Y) \end{aligned} \tag{9}$$

Integrating intervention effect and the objective function Eq. 6, the final objective is defined as below. The additional term $I(\bar{\mathbf{Z}},Y)$ is the key to evaluating the intervention effect.

$$L(\phi) = \min_{\phi} \underbrace{I(\mathbf{Z}; \mathbf{X}) - (I(\mathbf{Z}; Y)}_{\text{(1)positive term}} - \underbrace{\lambda I(\bar{\mathbf{Z}}; Y))}_{\text{(2)negative term}} \tag{10}$$

To intuitively understand the final objective Eq. 10, we divide it into positive and negative parts. The first positive term aims at finding minimal causes, and it helps retain information from the prediction task and drops redundant information from the original input. For the negative term, it is used to distinguish causes from all correlated variables by decreasing the information overlapping between Y and intervened representation $\bar{\mathbf{Z}}$.

5 PRACTICAL ALGORITHMS

In this section, we provide the details of how to evaluate the mutual information term in Eq. 10 and the alternative robust training process of our method.

5.1 Implementation of $L(\phi)$

In this paper, all objective functions are defined under mutual information formulation. We evaluate Eq.10 in two parts. The first positive part (Eq.10 (1)) is evaluated by the following parameterized objective, the variational estimation of mutual information [1]:

$$\begin{aligned} &I(\mathbf{Z};\mathbf{X}) - \lambda I(\mathbf{Z};Y) \\ &\geq \lambda \mathbb{E}_D[\mathbb{E}_{\mathbf{z}\in q(\mathbf{z}|\mathbf{x})}[\log p_g(y|\mathbf{z})] - \mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}))] \end{aligned} \tag{11}$$

For the negative term described in Eq. 10, the minimization process requires the upper bound of it [8]. The upper bound is formed as below:

$$I(\bar{\mathbf{Z}}, Y) \leq \mathbb{E}_{p(y,\bar{\mathbf{z}})}[\log(p(y|\bar{\mathbf{z}}))] - \mathbb{E}_y\mathbb{E}_{\bar{\mathbf{z}}}[\log(p(y|\bar{\mathbf{z}}))] \tag{12}$$

Note that the expectation on second term in Eq. 12 requires marginal distribution $p(y)\ p(\bar{\mathbf{z}})$ rather than joint distribution, therefore we independently sample y in practice. The intervened network (Fig. 2) helps us calculate the value of $\bar{\mathbf{Z}}$ along two steps. The first step, we build a neural network to generate the transformation vector $\mathbf{T} \in \mathbb{R}^t$ from observational data $\mathbf{X}$, where $(\mathbf{t} = \mathbf{k}(\mathbf{x}))$ and $\mathbf{k} : \mathcal{X} \to \mathbb{R}^t$ is a deterministic function modeled by neural network. The second step, the density of intervened $\bar{\mathbf{Z}}$ is calculated by $p(\bar{\mathbf{z}}|\mathbf{z},\mathbf{t}) = \delta(\bar{\mathbf{z}} = \mathbf{z} + \mathbf{t})$, where $\delta(\cdot)$ is Dirac delta function. In experiments, if $\mathbf{t}$ is close to 0, it will decline performance of our method since original $\mathbf{z}$ is close to the intervened one $\bar{\mathbf{z}}$. To avoid this problem, we add an additional constraint $\min_{\mathbf{k}} |\mathbf{t}^2 - \mathbf{b}|^2$, where b is a hyperparameter, in our experiments, we set $b = 0.8$.

5.2 Robust Learning under Adversarial Attack

To enhance the robustness against potential exogenous variable or noises $\boldsymbol{\epsilon}$ and guarantee the robustness of the proposed method, we extend our method by incorporating adversarial learning). Considering the causal generative process as $\mathbf{Y} = f(\mathbf{pa}_{\mathbf{Y}}, \boldsymbol{\epsilon}$, the $\boldsymbol{\epsilon}_1$ is regarded as a random noise perturbing the $\mathbf{pa}_{\mathbf{y}}$ inside a ball with finite diameter. We treat the inference approach as the process of adversarial attack [5, 6, 41] and define the ’Actions’-step in counterfactual estimation as

$$\begin{aligned} \mathbf{z}' &= \mathbf{z} + \boldsymbol{\epsilon}, \mathbf{z}' \in \mathcal{B}(\mathbf{z}, \beta) \\ \bar{\mathbf{z}}' &= \bar{\mathbf{z}} + \boldsymbol{\epsilon}, \bar{\mathbf{z}}' \in \mathcal{B}(\bar{\mathbf{z}}, \beta) \end{aligned} \tag{13}$$

where $\mathcal{B}(\mathbf{z}, \beta)$ is Wasserstein ball, in which the p-th Wasserstein distance [23] W_p [1]between z and $\mathbf{z}'$ is smaller than β. $\mathbf{z}'$ and $\bar{\mathbf{z}}'$ integrate both intervention and exogenous information. We further define intervention robustness (IR) to measure the worst intervention results of the intervention term in Eq.10. IV defines aims at finding the worst perturbation of $\mathbf{z}$ and $\bar{\mathbf{z}}$, which is formally defined below,

Definition 4. (Intervention Robustness) Let $\bar{\mathbf{Z}}'$ denote intervened variables on $\mathbf{Z} = \phi(\mathbf{X})$, $\forall \mathbf{z}' \in \mathcal{B}(\mathbf{z}, \beta), \bar{\mathbf{z}}' \in \mathcal{B}(\bar{\mathbf{z}}, \beta)$, D and D' denote datasets sample from $p(\mathbf{z}', \mathbf{y})$ and $p(\bar{\mathbf{z}}', \mathbf{y})$, the vulnerability of robust counterfactual estimation is defined as

$$\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} IR_{\mathcal{B}} = \min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)} I(Y;\mathbf{Z}') - \max_{\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} I(Y;\bar{\mathbf{Z}}') \tag{14}$$

Remark. The intervention robustness defines the worst intervention effect influenced by exogenous $\boldsymbol{\epsilon}$. For the representation $\mathbf{z}$, the term $\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)} I(Y;\mathbf{Z}')$ aims at find the perturbed $\mathbf{z}'$ around $\mathbf{z}$ with lowest mutual information $I(Y;\mathbf{Z}')$. For the transformed variable $\bar{\mathbf{z}}$, IV aims to find worst mutual information $\max_{\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} I(Y;\bar{\mathbf{Z}}')$. Combining two worst mutual information together, the IR term aims at finding the worst intervention effect perturbed by $\boldsymbol{\epsilon}$.

Combining the IR term with the original objective $L(\phi)$, we get the final objective function optimized by minmax approach. Equivalently, we only need to optimize $I(\mathbf{Z}';Y)$ rather than $I(\mathbf{Z};Y) + I(\mathbf{Z}';Y)$ since if the worst case $I(\mathbf{Z}';Y)$ is satisfied, $I(\mathbf{Z};Y)$ is satisfied. The robust optimization objective function is $L_{\mathrm{rb}}\phi$, where

$$\begin{aligned} &\max_{\phi} \min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} L(\phi) + IR_{\mathcal{B}} \\ &\geq \max_{\phi} \min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} \underbrace{I(\mathbf{Z}';Y) - \lambda I(\mathbf{Z};\mathbf{X})}_{\text{(1) positive}} - \underbrace{I(\bar{\mathbf{Z}}';Y)}_{\text{(2) negative}} \\ &= L_{\mathrm{rb}}(\phi) \end{aligned} \tag{15}$$

The inequality in above objective is due to

$$I(\mathbf{Z}';Y) - I(\bar{\mathbf{Z}}';Y) \geq \min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)} IR_{\mathcal{B}}.$$

The robust method is learned by the minimax procedure. Literally, the minimization procedure helps to avoid the worst-case led by exogenous variable $\boldsymbol{\epsilon}$, because it maximizes the intervention robustness by adjusting the parameter of feature extractor ϕ. The optimization objective of the robust method can extract minimal sufficient causal representation from observation data with high robustness ability.

We train and evaluate the robust method by the adversarial attack on representation space. We use PGD attack [18] with ∞-norm and 2-norm to get intervened $\mathbf{z}'$ and $\bar{\mathbf{z}}'$. We set $p_{\boldsymbol{\theta}}(\mathbf{z})$ as $\mathcal{N}(y, 1)$ to avoid trivial representations. Then we use negative cross entropy to approximate mutual information. More implementation details are shown in the supplementary material.

6 WHY CAUSAL REPRESENTATION CAN ENHANCE GENERALIZATION ABILITY

In this section, we theoretically analyze the generalization property of causal representation by learning theory framework [31]. Learning theory contains a set of methodologies to show the upper bound of the gap between risk/error on training data and all possible data from the data distribution. These methods justify a generalization problem that whether a model learned from a small data set can be generalized to any unseen test data from data distribution. Instead of estimating the IR risk bound, we start from the perspective of information theory and follow the framework of information bottleneck [32]. We provide a finite sample bound of the difference between ground truth and estimated one, which measures the generalization ability. The bound the relationship between $I(\mathbf{Z};Y)$ and its estimation $\hat{I}(\mathbf{Z};Y)$.

6.1 The Generalization Error Bound of i.i.d. Data

Here, we provide theoretical justification with the following theory (The proof is provided in supplementary material):

Theorem 6.1. *Let $\mathbf{Z} = \phi(\mathbf{X})$ where $\phi : \mathcal{X} \to \mathcal{Z}$ be a fixed arbitrary function, determined by a known conditional probability distribution $p(\mathbf{z}|\mathbf{x})$. Let m be sample size and C is a constant. For any*

[1] $\mathrm{W}_p(\mu, \nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathcal{Z}\times\mathcal{Z}} \Delta(z, z')^p\, d\gamma(z, z')\right)^{1/p}$, $\Gamma(\mu, \nu$ is the collection of all probability measures on $\mathcal{Z} \times \mathcal{Z}$

confidence parameter $0 < \delta < 1$*, it holds with a probability of at least* $1 - \delta$*, that*

1. General case (The learned representation Z *contains correlated information)*

$$|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\log(m) + \frac{1}{2}\sqrt{|\mathcal{Z}|}\log(|\mathcal{Y}|)\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{16}$$

where $m \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{Z}|\mathrm{e}^2$

2. Ideal case (The learned representation $\mathbf{Z}$ *contains information of causes)*

$$|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{\beta}\log(m) + \frac{1}{2}\sqrt{|\mathcal{Z}|}\log(|\mathcal{Y}|)\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}}$$

where $m \geq C\log(|\mathcal{Y}|/\delta)\beta\mathrm{e}^2$

Remark. The theorem provides a generalization bound under finite sample settings. It shows that when representation $\mathbf{Z}$ fully contains parent information $\mathbf{pa_Y}$, we achieve a sample complexity bound as $m \geq C\log(|\mathcal{Y}|/\delta)\beta\mathrm{e}^2$, where β is the variance of $\boldsymbol{\epsilon}$. The minimum number of samples needed reduces from $|\mathcal{Z}|$ to β, which is a tighter bound since in most of cases we assume $|\mathcal{Z}| \gg \beta$. This shows that $\mathbf{z} = \mathbf{pa_Y}$ gives the reduced sample complexity and tightened generalization bound. The theorem also serves as a general solution to causality prediction problems, supporting the claim that a better prediction is achieved with causal variables, compared to that with correlated variables.

6.2 The Generalization Error Bound when Distribution Shift Happens

We also show additional generalization results. For the scenario of distribution sift, we define the mutual information on source domain as $I_{\mathcal{S}}(\mathbf{Z}, Y)$ and mutual information on target domain as $I_{\mathcal{T}}(\mathbf{Z}, Y)$. Denote joint distribution in source and target domain as $\mathcal{S}(\mathbf{z}, y) = p_{\mathcal{S}}(\mathbf{z}, y)$ and $\mathcal{T}(\mathbf{z}, y) = p_{\mathcal{T}}(\mathbf{z}, y)$, separately.

The causal mechanism $p(\mathrm{y}|\mathbf{z})$ and causal representation $p(\mathbf{z})$ are stable under distribution shift such that $p_{\mathcal{S}}(\mathrm{y}|\mathbf{z} = \phi(\mathbf{x})) = p_{\mathcal{T}}(\mathrm{y}|\mathbf{z} = \phi(\mathbf{x}))$ and $p_{\mathcal{S}}(\mathbf{z} = \phi(\mathbf{x})) = p_{\mathcal{T}}(\mathbf{z} = \phi(\mathbf{x}))$, if $\mathbf{Z}$ is sufficient cause of Y.

Theorem 6.2. *Let* $\mathbf{Z} = \phi(\mathbf{X})$ *where* $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ *be a fixed arbitrary function, determined by a known conditional probability distribution* $p(\mathbf{z}|\mathbf{x})$*. Let* m *be sample size and* C *is a constant. In domain adaptation scenario, defining* $D_{KL}(\mathcal{S}||\mathcal{T}) > 0$ *as the Kullback-Leibler divergence between source domain and target domain. For any confidence parameter* $0 < \delta < 1$*, it holds with a probability of at least* $1 - \delta$*, that*

1. General case (The learned representation Z *contains correlated information)*

$$|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\log(m) + D_{KL}(\mathcal{T}||\mathcal{S}) + DI_{\mathcal{S}}\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{17}$$

2. Ideal case (The learned representation $\mathbf{Z}$ *contains information of sufficient causes of* Y*, Assumption 6.2 holds)*

$$|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\beta|}\log(m) + DI_{\mathcal{S}}\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{18}$$

where $D = \frac{2}{\min_{\mathbf{z}} p(\mathbf{z})}$ *and* $I_{\mathcal{S}} = \mathbb{E}_{\mathcal{S}(\mathbf{z},y)}\frac{\hat{p}(\mathbf{z},y)}{\hat{p}(\mathbf{z})\hat{p}(y)}$

Remark. The theorem shows that in a domain adaptation scenario, causal representation can help to achieve better generalization ability. We bound the risk of mutual information evaluation on the target domain by the bound on the source domain. It is because, in the training process, the information from the target domain is not observable. From the bounds of $|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})|$ shown in the general case and ideal case, we can see that the generalization error bound of the ideal case is smaller than that of the general case, with a margin quantified by a positive term $D_{KL}(\mathcal{S}||\mathcal{T}) > 0$. These theoretical results support that the causal representation can achieve better generalization ability under distribution shift.

7 EXPERIMENTS

In this section, we conduct extensive experiments to verify the effectiveness of our framework. In the following, we begin with the experiment setup, and then report and analyze the results.

7.1 Datasets

Our experiments are based on one synthetic and four real-world benchmarks. With the synthetic dataset, we evaluate our method in a controlled manner under the selected dataset. We follow the causal graph defined in Fig.1 (a) to build our synthetic simulator, on which we compare the representation learnt by our method with the ground truth under different β degrees.

7.2 Synthetic Datasets

The synthetic data is generated following the general causal graph Fig.1. We build the simulator using nonlinear functions refering to [43, 48]. We simulate 500 data for each settings. Let $\kappa_1(\cdot)$ and $\kappa_2(\cdot)$ as piecewise functions, and $\kappa_1(x) = x - 0.5$ if $x > 0$, otherwise $\kappa_1(x) = 0$, $\kappa_2(x) = x$ if $x > 0$, otherwise $\kappa_2(x) = 0$ and $\kappa_3(x) = x + 0.5$ if $x < 0$, otherwise $\kappa_3(x) = 0$. . For the fair evaluation, we set the same dimension for $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$ that $d_1 = d_2 = d_3 = 5$. The nonlinear systems are:

$$\begin{aligned}
&\mathbf{pa_Y} \sim U(-1, 1),\\
&\boldsymbol{\epsilon}_1 = \boldsymbol{\epsilon}_2 = \boldsymbol{\epsilon}_3 \sim \mathcal{N}(0.3, \beta I)\\
&\mathbf{nd}_1 = \boldsymbol{a}^T\kappa_1(\kappa_2([\mathbf{pa_Y}, \boldsymbol{\epsilon}_2])) + q,\\
&\mathbf{nd}_2 = \boldsymbol{a}^T\kappa_3(\kappa_2([-\mathbf{pa_Y}, -\boldsymbol{\epsilon}_2])) + q,\\
&\mathbf{nd_Y} = \sigma(\mathbf{nd}_1 + \mathbf{nd}_1 \cdot \mathbf{nd}_2)\\
&\mathbf{y}_1 = \boldsymbol{a}^T\kappa_1(\kappa_2([\mathbf{pa_Y}, \boldsymbol{\epsilon}_1])) + q,\\
&\mathbf{y}_2 = \boldsymbol{a}^T\kappa_3(\kappa_2([-\mathbf{pa_Y}, -\boldsymbol{\epsilon}_1])) + q,\\
&\mathbf{nd_Y} = \mathbb{I}(\sigma(\mathbf{y}_1 + \mathbf{y}_1 \cdot \mathbf{y}_2))\\
&\mathbf{dc}_1 = \boldsymbol{a}^T\kappa_1(\kappa_2([y, \boldsymbol{\epsilon}_3])) + q,\\
&\mathbf{dc}_2 = \boldsymbol{a}^T\kappa_3(\kappa_2([-y, -\boldsymbol{\epsilon}_3])) + q, \mathbf{dc_Y} = \sigma(\mathbf{dc}_1 + \mathbf{dc}_1 \cdot \mathbf{dc}_2)\\
&\mathbf{X} = [\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}]
\end{aligned}$$

Table 1: Overall Results on Yahoo!R3-OOD, Yahoo!R3-i.i.d. and PCIC

Dataset	Method	p=∞				p=2			
	Metrics	AUC	ACC	advAUC	advACC	AUC	ACC	advAUC	advACC
Yahoo!R3-OOD	base(robust)	0.5	0.4508	0.5	0.4508	0.5	0.4545	0.5	0.4537
	base(standard)	0.6198	0.6097	0.5212	0.5189	0.621	0.6099	0.5139	0.5188
	IB(standard)	0.6181	0.6063	0.5333	0.5149	0.6184	0.6069	0.5431	0.5255
	r-CVAE(robust)	0.6186	0.6235	0.5886	0.5912	0.6191	0.6241	0.5882	0.5907
	r-CVAE(standard)	0.6253	0.6249	0.5855	0.5863	0.6233	0.6243	0.5865	0.5872
	CaRI(robust)	0.6238	**0.6284**	**0.5993**	**0.5999**	0.6242	**0.6307**	**0.6008**	**0.601**
	CaRI(standard)	**0.629**	0.6257	0.5966	0.5965	**0.6276**	0.6255	0.5917	0.5917
Yahoo!R3-i.i.d.	base(robust)	0.5	0.6001	0.5	0.5997	0.5	0.6	0.5	0.6
	base(standard)	0.7334	0.7483	0.6267	0.6251	0.7346	0.752	0.6260	0.6103
	IB(standard)	0.7291	0.7513	0.6361	0.6721	0.7348	0.7521	0.6418	0.6775
	r-CVAE(robust)	0.7341	0.7093	0.7180	0.7080	0.7376	0.7151	0.7194	0.7082
	r-CVAE(standard)	0.7488	0.7515	0.7191	0.7072	0.7487	0.7529	0.7202	0.7099
	CaRI(robust)	0.7378	0.7168	**0.721**	**0.7107**	0.7374	0.7158	**0.7247**	**0.7159**
	CaRI(standard)	**0.7497**	**0.7503**	0.7191	0.7099	**0.7493**	**0.7495**	0.7188	0.7072
PCIC	base(robust)	0.5534	0.5875	0.5388	0.6257	0.5605	0.6498	0.5264	0.6287
	base(standard)	0.6177	0.6517	0.5231	0.589	0.6269	0.6615	0.519	0.5581
	IB(standard)	0.6242	0.6532	0.5741	0.6199	0.6216	0.6537	0.5768	0.6233
	r-CVAE(robust)	0.6363	0.6733	0.6088	0.6596	0.63	0.674	0.6187	0.6493
	r-CVAE(standard)	0.6358	0.6779	0.6138	0.6601	0.6328	0.6725	0.5893	0.6429
	CaRI(robust)	0.639	0.6761	**0.6225**	0.6638	0.6363	0.6709	**0.6332**	0.6576
	CaRI(standard)	**0.6447**	**0.6817**	0.6148	**0.664**	**0.6416**	**0.6803**	0.619	**0.6625**

where $q = 0.3$, $\mathbb{I}(x)$ is an indicator function, which is 1 if $x > 0$, and 0 otherwise. From synthetic data, we analyze whether CaRI has the ability to identify the $\mathbf{pa_Y}$ from mixed observational $\mathbf{X}$.

7.3 Real-word benchmarks

We also evaluate our method on real-world benchmarks for the recommendation system.

Yahoo! R3[2] is an online music recommendation dataset, which contains the user survey data and ratings for randomly selected songs. The dataset contains two parts: the uniform (OOD) set and the nonuniform (i.i.d.) set. The non-uniform (OOD) set contains samples of users deliberately selected and rates the songs by preference, which can be considered as a stochastic logging policy. For the uniform (i.i.d.) set, users were asked to rate 10 songs randomly selected by the system. The dataset contains 14,877 users and 1,000 items. The density degree is 0.812%, which means that the dataset only records 0.812% of rating pairs.

PCIC The dataset is collected from a survey by questionnaires about the rate and reason why the audience like or dislike the movie. Movie features are collected from movie review pages. The training data is a biased dataset consisting of 1000 users asked to rate the movies they care from 1720 movies. The validation and test set is the user preference on uniformly exposed movies. The density degree is set to be 0.241%.

For evaluation, Yahoo! R3 and Coat dataset both have two validation (include test) datasets. The i.i.d. set is 1/3 of data from a nonuniform logging policy, and the OOD set consists of the data generated under a uniform policy. For the PCIC dataset, we train our method on non-uniform datasets and perform evaluations on uniform datasets.

CelebA-anno The dataset contains more than 200K celebrity images, each with 40 attribute annotations. Following the previous work [14], we select 9 attribute annotations, which include Young, Male, Eyeglasses, Bald, Mustache, Smiling, Wearing Lipstick, and Mouth Open. Our task is to predict Smiling. $\mathbf{pa_Y}$ including {Young, Male}, $\mathbf{nd_Y}$ including {Eyeglasses, Bald, Mustache, Wearing Lipstick} and $\mathbf{cd_Y}$ including {Mouth Open}. From this dataset, we evaluate the ability to distinguish $\mathbf{pa_Y}$ from $\mathbf{X}$ (Results on CelebA-anno are provided in supplementary materials).

Coat Shopping Dataset[3] is a commonly used dataset collected from web-shop ratings on clothing. The self-selected ratings are the i.i.d. set and the uniformly selected ratings are the OOD set. In the training dataset, users were asked to rate 24 coats selected by themselves from 300 item sets. The test dataset collects the user rates on 16 random items from 300 item sets. Just as Yahoo! R3, the training dataset is a non-uniform dataset and the test dataset is a uniform dataset. The dataset provides side information on both users and item sets. The feature dimension of the user/item pair is 14/33.

Compared Method. For all the compared methods, we use the same model architecture, with different training strategies. The model consists of representation learning module $\mathbf{z} = \phi(\mathbf{x})$ and the downstream prediction module $\hat{\mathbf{y}} = g(\mathbf{z})$, with each module implemented by neural networks. **Base** model has no additional constraints on representation, and the optimization is to minimize the cross-entropy between y and learned $\hat{y}$. We involve a recently

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=r

[3] https://www.cs.cornell.edu/ schnabts/mnar/

Table 2: Overall Results on Coat dataset.

Dataset	Method	p=∞				p=2			
	Metrics	AUC	ACC	advAUC	advACC	AUC	ACC	advAUC	advACC
Coat-OOD	base(robust)	0.5586	0.5569	0.5479	0.5451	0.5593	0.556	0.5441	0.5412
	base(standard)	0.5659	0.5724	0.3874	0.4024	0.5642	0.5687	0.3128	0.3317
	IB(standard)	0.5659	0.5681	0.4701	0.4796	0.5659	0.5713	0.5442	0.5495
	r-CVAE(robust)	0.5629	0.5586	0.559	0.5544	0.5634	0.5591	0.5572	0.5522
	r-CVAE(standard)	0.5656	0.5643	0.5527	0.5478	0.5671	0.5649	0.5586	0.554
	CaRI(robust)	**0.5707**	0.5681	**0.5653**	**0.5659**	0.5705	0.5675	**0.5674**	**0.565**
	CaRI(standard)	0.5705	**0.5718**	0.5643	0.5659	**0.5725**	**0.5732**	0.5608	0.5601
Coat-i.i.d.	base(robust)	0.7156	0.7232	0.7034	0.7107	0.7195	0.7261	0.7001	0.7057
	base(standard)	0.7191	0.7217	0.4911	0.487	0.7235	0.7255	0.3642	0.3515
	IB(standard)	0.7162	0.72	0.6023	0.6017	0.7182	0.7222	0.694	0.696
	r-CVAE(robust)	0.7147	0.7222	0.7105	0.7181	0.7087	0.7169	0.7058	0.7141
	r-CVAE(standard)	0.7106	0.7184	0.7029	0.7106	0.7129	0.7206	0.7023	0.7059
	CaRI(robust)	0.7276	0.7339	**0.7208**	**0.727**	**0.7265**	**0.7331**	**0.7196**	**0.7261**
	CaRI(standard)	**0.7283**	**0.7355**	0.7125	0.7196	0.7248	0.7305	0.7069	0.7125

proposed variational estimation with information bottleneck (**IB**) [1], extend the condition VAE (CVAE [36]) by robust training process as **r-CVAE**, whose objective function is similar with CaRI but without a negative term (Eq.10 (2)). We conduct ablation studies by comparing our proposed method **CaRI** with the r-CVAE to evaluate the effectiveness of negative term. We evaluate our method on two main aspects: (i) **Generalization** of the model under distribution shifts and (ii) **Robustness** under adversarial attack on representation space. For (i), we evaluate our method on OOD and i.i.d. setting on Yahoo! R3 and Coat. For (ii), the standard mode of adversarial attack ($\beta = 0$) means that we do not perturb original $\mathbf{z}$. In robust mode, we set $\beta = \{0.1, 0.2, 0.1, 0.3, 0.3\}$ for PCIC, Yahoo! R3, Coat, Synthetic and CelebA-anno respectively.

Metrics. We use the common evaluation metrics AUC/ACC [11, 28] on CTR prediction and their variants called adv-ACC/ adv-AUC [18] on adversarially perturbed evaluation dataset. Moreover, we consider Distance Correlation metrics [13] to evaluate the similarity between learned representation and parental information $\mathbf{pa}_{\mathbf{Y}}$.

7.4 Implementation

Architecture and Setups: The model consists of two parts, the representation learning part and the downstream prediction part. For the representation learning part, we first use encode function $\phi(\cdot)$ to get representation $\mathbf{z}$ and get the intervened $\hat{\mathbf{z}}$. Then we perturb the learned $\mathbf{z}$ and $\bar{\mathbf{z}}$ by PGD attack [18] procedure to find the worst case corresponding to the worst downstream loss. We use PGD attack with ∞-norm ($p = \infty$) and 2-norm ($p = 2$) in our implementation. Finally we put $\mathbf{z}'$ and $\hat{\mathbf{z}}'$ into the downstream prediction model $g(\cdot)$ to calculate y. The likelihood in Eq.11 is estimated by cross-entropy loss. Note that the perturbation approach would block the gradient propagation between the representation learning process and downstream prediction in some implementation ways. Thus we use the conditional Gaussian prior $p_\theta(\mathbf{z}) = \mathcal{N}(y\mathbf{1}, \mathbf{I})$ rather than standard Gaussian distribution $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to calculate KL term. If gradient propagation is blocked, by using conditional prior, the learning process of representation $\mathbf{z}$ and exogenous $\boldsymbol{\epsilon}$ embedded in $\mathbf{z}'$ will not be influenced. The form of conditional Gaussian prior is more general $p_\theta(\mathbf{z}) = \mathcal{N}(\zeta(y), \mathbf{I})$, where $\zeta(\cdot)$ could be any non-trivial function like linear function even neural network.

Hyperparameter The hyper-parameters are determined by grid search. Specifically, the learning rate and batch size are tuned in the ranges of $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$ and $[64, 128, 256, 512, 1024]$, respectively. The weighting parameter λ is tuned in $[1, 100]$. Perturbation degrees are set to be $\beta = \{0.1, 0.2, 0.1, 0.3\}$ for Coat, Yahoo!R3, PCIC and CPC separately. The representation dimension is empirically set as 64. All the experiments are conducted based on a server with a 16-core CPU, 128g memories and an RTX 5000 GPU.

7.5 Overall Effectiveness

Table 1 shows the overall results on Yahoo! R3 and PCIC. From Yahoo! R3 dataset, which contains both i.i.d. and OOD validation and test sets, we find that our method enjoys better generalization. In Yahoo! R3 OOD, our method increases the performance by 1.9%, and 8.1%, in terms of ACC and adv-ACC, compared with the base method. The performance of r-CVAE is close to CaRI, since it is a modified version of our method, which only includes the positive term in Eq.15 but removes the negative term. The difference between the performances of CaRI and r-CVAE shows the effectiveness of the negative term in the objective function of CaRI. In PCIC dataset, standard and robust modes of CaRI achieve the best AUC at 64.47%, and 63.9% respectively, which validates the effectiveness of our idea. In the robust training mode, our method achieves the best performance in adversarial metrics. In the PCIC dataset, our method reaches 62.25%, which increases by 8.37% against the base method on adv-AUC. Robust training of CaRI is also better than the standard training, winning with a margin of around 1.42%. we present the additional test results and analysis in this section. Table 2 shows overall experimental results on Coat, The table contains

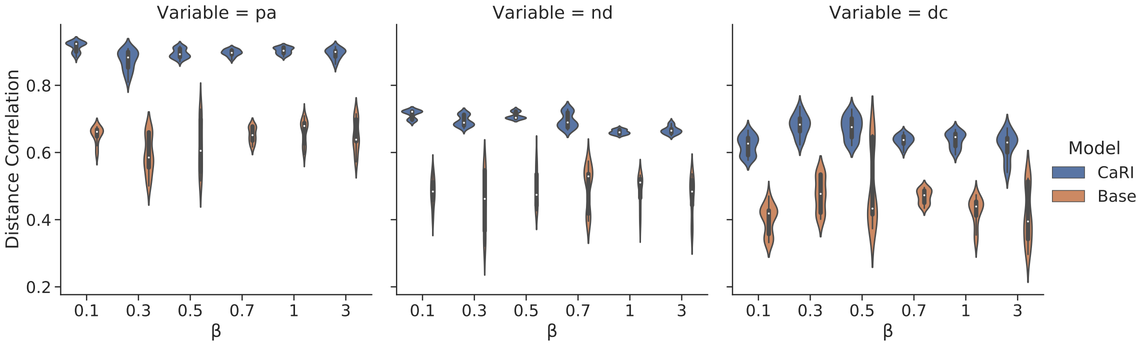


Figure 2: Representation learning results on synthetic dataset over different range of β, where $p = 2$ under robust training.

both i.i.d. and OOD settings. Based on this we find that in most cases, our method achieves better performance in terms of AUC and ACC, compared to base methods. The overall results show that the robust learning process with exogenous variables involved enhances the adversarial performance on perturbed samples. On the other hand, in standard training mode, CaRI achieves better adversarial performance than baselines including base method and IB. We find that standard training of CaRI on PCIC has an AUC of 64.47%, which is better than the performance under robust training (63.9%). But contrary conclusions are drawn on adversarial performance. The result supports that the causal representation we learned is more robust. The performance of the base method in robust training mode is worst in most of cases, indicating that the robust training process will largely influence the learning of the model and ruin the prediction model. Although the robust training deteriorates the performance of on normal dataset, it will help to identify the causal representation, which benefits downstream prediction under adversarial attack. The robust learning process with exogenous variables involved enhances the adversarial performance on perturbed samples. On the other hand, in standard training mode, CaRI achieves better adversarial performance than all baselines. The result supports that the causal representation our method learned is more robust.

7.6 Representation Analysis

In this section, we study whether our method CaRI helps to identify the parental information from observational data. Fig.2 demonstrates the ability of the model to learn causal representations under different β degrees on a synthetic dataset. The figure shows the distance correlation between the learned representation $\mathbf{z}$ and different parts of observational data, namely ($\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$). From Fig.2 (left), we find that our method learns a representation that is with the highest similarity, in comparison with the base method under different values of β. It is evidence that our method successfully identifies the parental information from mixed observational data. The information from $\mathbf{nd_Y}$ and $\mathbf{dc_Y}$ are not considered as important as parental information from CaRI, and the distance correlation metric corresponding to this part is slightly lower. We also find that the metric under CaRI gets lower variance, which shows the stable performance of CaRI. On the contrary, the distance correlation metric of the base method is with high variance, which indicates the possible incapability of the base method on extracting the parental information from observations.

8 CONCLUSIONS & FUTURE WORKS

In this paper, we deal with the problem of learning causal representations from observational data, which comes with satisfactory generalization ability. Assuming that the underlying latent factors follow some causal generative models, we argue that learning a minimum sufficient cause of the system is the optimal solution. By analyzing the information theoretical property of our hypothetical graphical model, we propose a causality-inspired representation learning method by optimizing a function with regularized mutual information constraints. It achieves effective learning with guaranteed sample complexity reduction under certain assumptions. Extensive experiments on real-world dataset show the effectiveness of our algorithm, verifying our claim of robustness of the representation with respect to downstream tasks.

Future Works. For future works, one promising direction is to involve the concept of Kolmogorov complexity in information theory. Different to mutual information and information entropy, Kolmogorov complexity is an asymmetric notion. Based on such a concept, we can develop a causal representation learning method without introducing an intervention network. Another direction is that our proposed method can be generalized to a mixture of anti-causal and causal learning frameworks where observation data contains both parents and descendants of outcome label Y. The information-theoretic-based sample complexity theorem can inspire the generalization error/risk analysis on causal representation learning and causal structure learning. Lastly, this paper is based on the assumption of the given causal graph Fig. 1. In the future, it is interesting to extend our method to more complex scenarios like sequential prediction, reinforcement learning etc.

9 ACKNOWLEDGEMENTS

The work was supported by National Key R&D Program of China (2022YFB4501500, 2022YFB4501504). We thank Jianhong Wang for the fruitful discussion on some of the results and for proofreading this paper and thank Xu Chen for proofreading. We thank the anonymous reviewers for their feedback.

REFERENCES

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[3] Nihat Ay and Daniel Polani. 2008. Information Flows in Causal Networks. *Adv. Complex Syst.* 11, 1 (2008), 17–41. https://doi.org/10.1142/S0219525908001465

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research)*, Jennifer G. Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 530–539. http://proceedings.mlr.press/v80/belghazi18a.html

[5] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. 2009. *Robust optimization.* Princeton university press.

[6] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018), 317–331.

[7] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research)*, Vol. 119. PMLR, 1779–1788. http://proceedings.mlr.press/v119/cheng20b.html

[8] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*. PMLR, 1779–1788.

[9] Thomas M Cover. 1999. *Elements of information theory.* John Wiley & Sons.

[10] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. 2016. Domain adaptation with conditional transferable components. In *International conference on machine learning*. PMLR, 2839–2848.

[11] Asela Gunawardana and Guy Shani. 2009. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research* 10, 12 (2009).

[12] Dominik Janzing and Bernhard Schölkopf. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory* 56, 10 (2010), 5168–5194.

[13] Terry Jones, Stephanie Forrest, et al. 1995. Fitness Distance Correlation as a Measure of Problem Difficulty for Genetic Algorithms.. In *ICGA*, Vol. 95. 184–192.

[14] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. 2017. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023* (2017).

[15] Erich Leo Lehmann and Henry Scheffé. 2012. Completeness, similar regions, and unbiased estimation-Part I. In *Selected Works of EL Lehmann*. Springer, 233–268.

[16] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[17] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. 2021. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353* (2021).

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[19] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. 2017. Domain adaptation by using causal inference to predict invariant conditional distributions. *arXiv preprint arXiv:1707.06422* (2017).

[20] Nicolai Meinshausen. 2018. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*. IEEE, 6–10.

[21] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17, 1 (2016), 1103–1204.

[22] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1933–1942.

[23] Victor M Panaretos and Yoav Zemel. 2019. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application* 6 (2019), 405–431.

[24] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. 2018. Learning independent causal mechanisms. In *International Conference on Machine Learning*. PMLR, 4036–4044.

[25] Judea Pearl. 2009. *Causality.* Cambridge university press.

[26] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (2016), 947–1012.

[27] Aahlad Puli, Adler Perotte, and Rajesh Ranganath. 2020. Causal estimation with functional confounders. *Advances in neural information processing systems* 33 (2020), 5115–5125.

[28] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[29] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 2018. Invariant models for causal transfer learning. *The Journal of Machine Learning Research* 19, 1 (2018), 1309–1342.

[30] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[31] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms.* Cambridge university press.

[32] Ohad Shamir, Sivan Sabato, and Naftali Tishby. 2010. Learning and generalization with the information bottleneck. *Theoretical Computer Science* 411, 29-30 (2010), 2696–2711.

[33] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. 2020. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637* (2020).

[34] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards Out-Of-Distribution Generalization: A Survey. *arXiv preprint arXiv:2108.13624* (2021).

[35] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. 2018. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering* 31, 2 (2018), 357–370.

[36] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015), 3483–3491.

[37] Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. 2021. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In *International Conference on Machine Learning*. PMLR, 9848–9858.

[38] Bastian Steudel, Dominik Janzing, and Bernhard Schölkopf. 2010. Causal Markov condition for submodular information measures. *arXiv preprint arXiv:1002.4020* (2010).

[39] Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, and Chi Xu. 2018. Recurrent knowledge graph embedding for effective recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 297–305.

[40] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*. PMLR, 6056–6065.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[42] Yixin Wang and Michael I Jordan. 2021. Desiderata for Representation Learning: A Causal Perspective. *arXiv preprint arXiv:2109.03795* (2021).

[43] Mengyue Yang, Quanyu Dai, Zhenhua Dong, Xu Chen, Xiuqiang He, and Jun Wang. 2021. Top-N Recommendation with Counterfactual User Preference Simulation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2342–2351.

[44] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. CausalVAE: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9593–9602.

[45] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. 2015. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*.

[46] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1449–1458.

[47] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503* (2021).

[48] Hao Zou, Kun Kuang, Boqi Chen, Peixuan Chen, and Peng Cui. 2019. Focused Context Balancing for Robust Offline Policy Evaluation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 696–704. https://doi.org/10.1145/3292500.3330852