IDENTIFYING TREATMENT EFFECTS UNDER UNOBSERVED CONFOUNDING BY CAUSAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As an important problem of causal inference, we discuss the estimation of treatment effects under the existence of unobserved confounding. By representing the confounder as a latent variable, we propose Counterfactual VAE, a new variant of variational autoencoder, based on recent advances in identifiability of representation learning. Combining the identifiability and classical identification results of causal inference, under mild assumptions on the generative model and with small noise on the outcome, we theoretically show that the confounder is identifiable up to an affine transformation and then the treatment effects can be identified. Experiments on synthetic and semi-synthetic datasets demonstrate that our method matches the state-of-the-art, even under settings violating our formal assumptions.

1 INTRODUCTION

Causal inference (Imbens & Rubin, 2015; Pearl, 2009), i.e, estimating causal effects of interventions, is a fundamental problem across many domains. In this work, we focus on the estimation of treatment effects, e.g., effects of public policies or clinical trials, based on a set of observations consisting of binary labels for treatment / control (non-treated), outcome, and other covariates. The fundamental difficulty of causal inference is that we never have observations of *counterfactual* outcomes, which would have been if we had made another decision (treatment or control). While the ideal protocol for causal inference is randomized controlled trials (RCTs), they often have ethical and practical issues, or are prohibitively expensive. Thus, causal inference from observational data is indispensable, though they introduce other challenges. Perhaps the most crucial one is *confounding*: there might be variables (called *confounders*) that causally affect both the treatment and the outcome, and spurious correlation follows.

Most of works in causal inference rely on the *unconfoundedness* assumption that appropriate covariates are collected so that the confounding can be controlled by conditioning on or *adjusting* for those variables. This is still challenging, due to systematic difference of the distributions of the covariates between the treatment and control groups. One classical way of dealing with this difference is re-weighting (Horvitz & Thompson, 1952). There are semi-parametric methods, which have better finite sample performance, e.g. TMLE (Van der Laan & Rose, 2011), and also non-parametric, tree-based, methods, e.g. Causal Forests (CF) (Wager & Athey, 2018). Notably, there is a recent rise of interest in *representation learning* for causal inference starting from Johansson et al. (2016).

There are a few lines of works that challenge the difficult but important problem of causal inference under *unobserved confounding*. Without covariates we can adjust for, many of them assume special structures among the variables, such as instrumental variables (IVs) (Angrist et al., 1996), proxy variables (Miao et al., 2018), network structure (Ogburn, 2018), and multiple causes (Wang & Blei, 2019). Among them, instrumental variables and proxy (or surrogate) variables are most commonly exploited. *Instrumental variables* are not affected by unobserved confounders, influencing the outcome only through the treatment. On the other hand, *proxy variables* are causally connected to unobserved confounders, but are not confounding the treatment and outcome by themselves. Other methods use restrictive parametric models (Allman et al., 2009), or only give interval estimation (Manski, 2009; Kallus et al., 2019).

In this work, we address the problem of estimating treatment effects under unobserved confounding. We further discuss the *individual-level* treatment effect, which measures the treatment effect conditioned on the covariate, for example, on a patient's personal data. To model the problem, we regard the covariate as a proxy variable and the confounder as a latent variable in representation learning.

Our method particularly exploits the recent advance of *identifiability* of representation learning for VAE (Khemakhem et al., 2020). The hallmark of deep neural networks (NNs) might be that they can learn representations of data. It is desirable that the learned representations are *interpretable*, that is, in approximately the same relationship to the latent sources for each down-stream task. A principled approach to this is identifiability, that is, when optimizing our learning objective w.r.t. the representation function, only a unique (maybe up to certain trivial transformation) optimum will be returned. Our method builds on this and further provides the stronger identifiability of representations that is needed in causal inference.

The proposed method is also based firmly on the well-established results in causal inference. In many work exploiting proxies, it is assumed that the proxies are independent of the outcome given the confounder (Greenland, 1980; Rothman et al., 2008; Kuroki & Pearl, 2014). Except this independence, we do not make any specific causal assumptions on the covariates. Further, our method naturally combines a new VAE architecture with the classical result of Rosenbaum & Rubin (1983) regarding the sufficient information for identification of treatment effects, showing identifiability proof of both latent representations and treatment effects.

The main **contributions** of this paper are as follows: 1) interpretable, causal representation learning by a new VAE architecture for estimating treatment effects under unobserved confounding; 2) theoretical analysis of the identifiability of representation and treatment effect; 3) experimental study on diverse settings showing performance of state-of-the-art.

2 Related work

Identifiability of representation learning. With recent advances in nonlinear ICA, identifiability of representations is proved under a number of settings, e.g., auxiliary task for representation learning (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019) and VAE (Khemakhem et al., 2020). Recently, Roeder et al. (2020) extends the the result to include a wide class of state-of-the-art deep discriminative models. The results are exploited in bivariate causal discovery (Monti et al., 2019; Wu & Fukumizu, 2020) and structure learning (Yang et al., 2020). To the best of our knowledge, this work is the first to explore this new possibility in causal inference.

Representation learning for causal inference. Recently, researchers start to design representation learning methods for causal inference, but mostly limited to unconfounded settings. Some methods focus on learning a balanced covariate representation, e.g., BLR/BNN (Johansson et al., 2016), and TARnet/CFR (Shalit et al., 2017). Adding to this, Yao et al. (2018) also exploits the local similarity of between data points. Shi et al. (2019) uses similar architecture to TARnet, considering the importance of treatment probability. There are also methods using GAN (Yoon et al., 2018, GANITE) and Gaussian process (Alaa & van der Schaar, 2017). Our method adds to these by also tackling the harder problem of unobserved confounding.

Causal inference with auxiliary structures. Both our method and CEVAE (Louizos et al., 2017) are motivated by exploiting proxies and use VAE as a learning method. However, CEVAE assumes a specific causal graph where the covariates should be independent of the treatment given the confounder. Further, CEVAE relies on the assumption that VAE can recover the true latent distribution. Miao et al. (2018) established conditions for identification using more general proxies, but without practical estimation method. Note that, two active lines of works in machine learning exist in their own right, exploiting IV (Hartford et al., 2017) and network structure (Veitch et al., 2019).

3 Setup and preliminaries

3.1 Treatment effects and confounders

Following Imbens & Rubin (2015), we begin by introducing *potential outcomes* (or *counterfactual outcomes*) $\mathrm{y}(t), t = 0, 1$. $\mathrm{y}(t)$ is the outcome we *would* observe, if we apply treatment value t. Note

that, for a unit under research, we can observe only one of $y(0)$ or $y(1)$, corresponding to which factual treatment we have applied. This is the *fundamental problem of causal inference*.

We write expected potential outcomes, conditioned on covariate(s) $\mathbf{x} = \boldsymbol{x}$ as $\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathrm{y}(t)|\mathbf{x} = \boldsymbol{x})$. The estimands in this work are the casual effects, which are Conditional Average Treatment Effect (CATE) and Average Treatment Effect (ATE) defined by

$$\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \qquad ATE = \mathbb{E}(\tau(\boldsymbol{x})) \tag{1}$$

CATE can be understood is an *individual-level* treatment effect conditioned on the covariate.

In general, we need three assumptions for identification (Rubin, 2005). There should exist variable $\mathbf{z} \in \mathbb{R}^n$ satisfies *ignorability* ($\mathrm{y}(0), \mathrm{y}(1) \perp\!\!\!\perp \mathrm{t}|\mathbf{z}$) and *positivity* ($\forall \boldsymbol{z}, t : p(\mathrm{t} = t|\mathbf{z} = \boldsymbol{z}) > 0$), and also given the *consistency* of counterfactuals ($\mathrm{y} = \mathrm{y}(t)$ if $\mathrm{t} = t$) (See Appendix for explanations). Then, treatment effects can be *identified* by:

$$\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\mathrm{y}(t)|\mathbf{z}, \mathbf{x} = \boldsymbol{x})) = \mathbb{E}(\mathbb{E}(\mathrm{y}|\mathbf{z}, \mathbf{x} = \boldsymbol{x}, \mathrm{t} = t)) = \int (\int p(y|\boldsymbol{z}, \boldsymbol{x}, t) y dy) p(\boldsymbol{z}|\boldsymbol{x}, t) d\boldsymbol{z} \tag{2}$$

The second equality uses the three conditions. We say that *strong ignorability* holds when we have both ignorability and positivity. In this work, we consider *unobserved confounding*, that is, we assume the existence of *confounder(s)* $\mathbf{z}$, satisfying the three conditions, but it is unobserved.

The following theorem adapted from Rosenbaum & Rubin (1983) is central to causal inference and we will use it for motivating and justifying our method. Such function $\boldsymbol{b}(\mathbf{z})$ is called a *balancing score* (of $\mathbf{z}$). Obviously, the *propensity score* $e(\mathbf{z}) := p(\mathbf{t} = 1|\mathbf{z})$, the propensity of assigning the treatment given $\mathbf{z}$, is a balancing score (with f be the identity function).

Theorem 1 (Balancing score)**.** *Let $\boldsymbol{b}(\mathbf{z})$ be a function of random variable $\mathbf{z}$. Then $\mathrm{t} \perp\!\!\!\perp \mathbf{z}|\boldsymbol{b}(\mathbf{z})$ if and only if $f(\boldsymbol{b}(\mathbf{z})) = p(\mathrm{t} = 1|\mathbf{z}) := e(\mathbf{z})$ for some function f (or more formally, $e(\mathbf{z})$ is $\boldsymbol{b}(\mathbf{z})$-measurable). Assume further that $\mathbf{z}$ satisfies strong ignorability, then so does $\boldsymbol{b}(\mathbf{z})$.*

3.2 Variational autoencoders

Variational autoencoders (VAEs) (Kingma et al., 2019) are a class of latent variable models with latent variable $\mathbf{z}$, and observed variable $\mathbf{y}$ is generated by distribution $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{z})$ (the decoder). The variational lower bound of the log-likelihood is written as:

$$\log p(\mathbf{y}) \geq \log p(\mathbf{y}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z}|\mathbf{y})) = \underbrace{\mathbb{E}_{\boldsymbol{z} \sim q} \log p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z}))}_{\mathcal{L}_{VAE}(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\phi})}, \tag{3}$$

where the encoder $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{y})$ is introduced to approximate the true posterior $p(\mathbf{z}|\mathbf{y})$. The decoder $p_{\boldsymbol{\theta}}$ and encoder $q_{\boldsymbol{\phi}}$ are usually parametrized by NNs. We will omit the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ in notations when appropriate. Using the reparameterization trick (Kingma & Welling, 2014) and optimizing the evidence lower bound (ELBO) $\mathbb{E}_{\mathbf{y} \sim \mathcal{D}}(\mathcal{L}(\boldsymbol{y}))$ with data $\mathcal{D}$, we train the VAE efficiently. Conditional VAE (CVAE) adds a conditioning variable $\mathbf{c}$ to (3) (See Appendix for details).

As mentioned, identifiable VAE (iVAE) (Khemakhem et al., 2020) provides the first identifiability result for VAE, using auxiliary variable $\mathbf{u}$. It also assumes $\mathbf{y} \perp\!\!\!\perp \mathbf{u}|\mathbf{z}$, that is, $p(\mathbf{y}|\mathbf{z}, \mathbf{u}) = p(\mathbf{y}|\mathbf{z})$. The variational lower bound is

$$\log p(\mathbf{y}|\mathbf{u}) \geq \underbrace{\mathbb{E}_{\boldsymbol{z} \sim q} \log p(\mathbf{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}, \mathbf{u}) \| p(\mathbf{z}|\mathbf{u}))}_{\mathcal{L}_{iVAE}(\mathbf{y},\mathbf{u})} \tag{4}$$

Note that, unlike CVAE, the decoder does *not* depend on $\mathbf{u}$ due to the independence assumption.

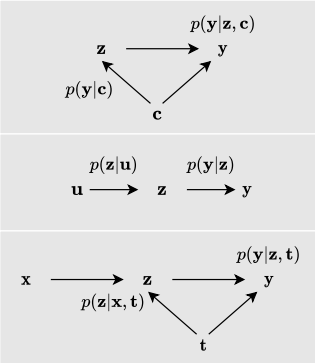

Figure 1: Graphical models of the decoders. From top: CVAE, iVAE, and CFVAE. The encoders for the VAEs are similar: they take all observed variables and build approximate posteriors.

4 VAE architecture for casual representation learning

The probabilistic model of our VAE follows naturally from the right most side of (2), where the involved distribution is $p(\mathrm{y}, \mathbf{z}|\mathbf{x}, \mathrm{t}) = p(\mathrm{y}|\mathbf{z}, \mathbf{x}, \mathrm{t}) p(\mathbf{z}|\mathbf{x}, \mathrm{t})$. If we treat covariate $\mathbf{x}$ as a proxy and assume further the conditional independence $\mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$ as in most work exploiting proxies, we have

$$p(\mathrm{y}, \mathbf{z}|\mathbf{x}, \mathrm{t}) = p(\mathrm{y}|\mathbf{z}, \mathrm{t}) p(\mathbf{z}|\mathbf{x}, \mathrm{t}) \tag{5}$$

A natural next step is to design a VAE for this joint distribution, which learns to recover a causal representation of $\mathbf{z}$. By "causal", we mean the representation can be used to identify or estimate treatment effects. Recovering the true confounder $\mathbf{z}$ would be great, but this is not required, as shown in Theorem 1, which says, for identification of treatment effects, we only need to have $\boldsymbol{b}(\mathbf{z})$, a causal representation of $\mathbf{z}$, which contains the information of propensity score $e(\mathbf{z})$, the part of $\mathbf{z}$ that is relevant to treatment assignment.

We are a step away from the VAE architecture now. Note that (5) has similar factorization with iVAE: $p(\mathbf{y}, \mathbf{z}|\mathbf{u}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\mathbf{u})$ from $\mathbf{y} \perp\!\!\!\perp \mathbf{u}|\mathbf{z}$, meaning we can use our covariate $\mathbf{x}$ as auxiliary variable $\mathbf{u}$ in iVAE from the independence assumption of proxy variable. Further from the conditioning on t in (5), we design a VAE architecture as a combination of CVAE and iVAE, with treatment t and covariate $\mathbf{x}$ as conditioning and auxiliary variable, respectively. The ELBO can be derived as

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}, \mathrm{t}) &\geq \log p(\mathbf{y}|\mathbf{x}, \mathrm{t}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \| p(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})) \\ &= \mathbb{E}_{\boldsymbol{z} \sim q} \log p(\mathbf{y}|\boldsymbol{z}, \mathrm{t}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \| p(\mathbf{z}|\mathbf{x}, \mathrm{t})) := \mathcal{L}_{CFVAE}(\mathbf{x}, \mathbf{y}, \mathrm{t}). \end{aligned} \tag{6}$$

As in iVAE, the decoder drops the dependence on $\mathbf{x}$. We name this architecture the *Counterfactual VAE* (CFVAE). Figure 1 depicts the relationship of CVAE, iVAE, and CFVAE.

We detail parameterization of CFVAE. The decoder $p_{\boldsymbol{f},\boldsymbol{g}}(\mathbf{y}|\mathbf{z}, \mathrm{t})$, conditional prior $p_{\boldsymbol{h},\boldsymbol{k}}(\mathbf{z}|\mathbf{x}, \mathrm{t})$, and encoder $q_{\boldsymbol{r},\boldsymbol{s}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})$ are factorized Gaussians, i.e., a product of 1-dimensional Gaussian distributions. This is not restrictive if the mean and variance are given by arbitrary nonlinear functions.

$$\mathbf{y}|\mathbf{z}, \mathrm{t} \sim \textstyle\prod_{j=1}^{d} \mathcal{N}(y_j; f_j, g_j), \quad \mathbf{z}|\mathbf{x}, \mathrm{t} \sim \prod_{i=1}^{n} \mathcal{N}(z_i; h_i, k_i), \quad \mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t} \sim \prod_{i=1}^{n} \mathcal{N}(z_i; r_i, s_i). \tag{7}$$

$\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{k})$ and $\boldsymbol{\phi} = (\boldsymbol{r}, \boldsymbol{s})$ are functional parameters given by NNs which take the respective conditional variables as inputs (e.g. $\boldsymbol{h} := (h_i(\boldsymbol{x}, t))^T$).

5 Identifiability of representation and treatment effects

In the following, we will show that CFVAE can identify the latent variable up to an affine transformation (Sec. 5.1), it can learn a balancing score as a causal representation, and its decoder is a valid estimator for potential outcomes (Sec. 5.2).

5.1 Identifiability of representation

In this subsection, we show that CFVAE can identify latent variable $\mathbf{z}$ up to an element-wise affine transformation when the noise on the outcome is small. Based on this result, we can gain insight on how to make CFVAE learn a balancing score.

With the learning model specified by (5) and (7)[1], as of Theorem 1 in Khemakhem et al. (2020), for $t = 0, 1$, assume that $\boldsymbol{f}_t(\boldsymbol{z}) := (f_i(\boldsymbol{z}, t))^T$ is injective, $\boldsymbol{g}_t(\boldsymbol{z}) = \boldsymbol{\sigma}_{\mathbf{y},t}$ is constant (i.e. $g_i(\boldsymbol{z}, t) = \sigma_{\mathrm{y}_i,t}$), and $\boldsymbol{\lambda}_t(\mathbf{x}) := (\boldsymbol{h}(\mathbf{x}, t), \boldsymbol{k}(\mathbf{x}, t))^T$, which is seen as a random variable, is not degenerate. Then, given $\mathrm{t} = t$, the model $p_{\boldsymbol{\theta}_t}(\mathbf{y}|\mathbf{x}, \mathrm{t} = t) = \int p_{\boldsymbol{\theta}_t}(\mathbf{y}, \mathbf{z}|\mathbf{x}, \mathrm{t} = t) d\mathbf{z}$ is identifiable up to an equivalence class, that is, for $t = 0, 1$, if $p_{\boldsymbol{\theta}_t} = p_{\boldsymbol{\theta}'_t}$[2], we have

$$\boldsymbol{f}_t^{-1}(\mathbf{y}_t) = \boldsymbol{A}_t \boldsymbol{f}_t'^{-1}(\mathbf{y}_t) + \boldsymbol{b}_t := \mathcal{A}_t(\boldsymbol{f}_t'^{-1}(\mathbf{y}_t)) \tag{8}$$

where $p(\mathbf{y}_t) := p(\mathbf{y}|t)$, $\boldsymbol{A}_t$ is an invertible n-square matrix and $\boldsymbol{b}_t$ is a n-vector. Similarly to Sorrenson et al. (2019), we can further show that $\boldsymbol{A}_t = \mathrm{diag}(\boldsymbol{a}_t)$ is a diagonal matrix. By a slight abuse of symbol, we will overload $|$ to make a shorthand for equations like (8), e.g., (8) can be written as $\boldsymbol{f}^{-1}(\mathbf{y}) = \mathcal{A}(\boldsymbol{f}'^{-1}(\mathbf{y}))|t$. Note that, by definition of inverse, we also have $\boldsymbol{f}' = \boldsymbol{f} \circ \mathcal{A}|t$.

The importance of model identifiability can be seen more clearly in the limit of small noise on $\mathbf{y}$. Corollary 1 can be easily understood by noting that after learning with small noise on $\mathbf{y}$, the encoder and decoder both degenerate to deterministic functions: in (7) $\boldsymbol{g}' = \boldsymbol{s}' = \mathbf{0}$, and $\forall \boldsymbol{x}, \boldsymbol{z}'_t = \boldsymbol{r}'_t(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{f}_t'^{-1}(\boldsymbol{y})$. Note that, we only assume the VAE learns *observational* distributions $p_{\boldsymbol{\theta}'_t}(\mathbf{y}|\mathbf{x}, \mathrm{t} = t)$ the same as the truth, but this leaves room for *latent* distributions different to the truth.

[1] We specified factorized Gaussians in (7) and they show good performance in our experiments. But Corollary 1 and Theorem 2 can be extended to more general exponential families, see Khemakhem et al. (2020).

[2] $\boldsymbol{\theta}'_t = (\boldsymbol{f}'_t, \boldsymbol{h}'_t, \boldsymbol{k}'_t)$ is another set of parameters giving the same distribution, which is learned by VAE. In this paper, symbol "$'$" (prime) always indicates parameters (variables, etc.) learned/recovered by VAE.

Corollary 1 (Identifiability of representation)**.** *For $t = 0, 1$, in the limit of $\boldsymbol{\sigma}_{\mathbf{y},t} \to \mathbf{0}$, assume CFVAE can learn a distribution $p_{\boldsymbol{\theta}'_t} = p_{\boldsymbol{\theta}_t}$, then the latent variable $\mathbf{z}$ and the mean parameter $\boldsymbol{f}_t$ of $\mathbf{y}$ can be identified up to an element-wise affine transformation: $\mathbf{z} = \mathcal{A}(\mathbf{z}')|t$ and $\boldsymbol{f}' = \boldsymbol{f} \circ \mathcal{A}|t$.*

This is a strong result for learning interpretable representation, but it is not enough for causal inference. To see this in a principled way, we recall the concept of balancing score. The recovered latent $\mathbf{z}'$ in Corollary 1 is not a balancing score, due to the different $\mathcal{A}_t$ for $t = 0, 1$. If $\mathbf{z}'$ *were* a balancing score, we would have $\mathrm{t} \perp\!\!\!\perp \mathbf{z}|\mathbf{z}'$. However, given $\mathbf{z}' = \boldsymbol{z}'$, $\boldsymbol{z} = \mathrm{diag}(\boldsymbol{a}_t)\boldsymbol{z}' + \boldsymbol{b}_t$ is a deterministic function of t, contradicting with $\mathrm{t} \perp\!\!\!\perp \mathbf{z}|\mathbf{z}'$. (A more concrete analysis can be found in Appendix.) This example also suggests that we will have a balancing score if we can get rid of the dependence on $\mathrm{t} = t$. The next subsection discusses some assumptions on $\mathbf{x}$ to remove the "$|t$" in Corollary 1.

5.2 Identifiability of treatment effects

The following definition will be used in Theorem 2. The importance of this definition is immediate from Theorem 1, that is, if a balancing *covariate* is also a function of $\mathbf{z}$, then it is a balancing *score*.

Definition 1 (Balancing covariate)**.** Random variable $\mathbf{x}$ is a *balancing covariate* of random variable $\mathbf{z}$ if $\mathrm{t} \perp\!\!\!\perp \mathbf{z}|\mathbf{x}$. We also simply say $\mathbf{x}$ is *balancing* (or *non*-balancing if it does not satisfy this definition).

Given that a balancing score of the true confounder is *sufficient* for strong ignorability, a natural and interesting question is that, does a balancing covariate of the true confounder also satisfies strong ignorability? The answer is *no*. To see why, and also to better understand the significance of Theorem 2, we provide a theoretical result (Proposition 1) indicating that a balancing covariate of the true confounder might *not* even satisfy *ignorability*. We also refer readers to Appendix 8.5 where we examine two definitions as important special cases of balancing covariate, one of those is noiseless proxy, which might *not* satisfy *positivity*.

Proposition 1. *Let $\mathbf{x}$ be a balancing covariate of $\mathbf{z}$. If $\mathbf{z}$ satisfies ignorability* and $\mathrm{y}(0), \mathrm{y}(1) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$, *then $\mathbf{x}$ satisfies ignorability.*

Given this proposition, we know our assumptions are weaker than ignorability, and our method can work under unobserved confounding ($\mathbf{x}$ might not satisfy ignorability). Note the independence $\mathrm{y}(0), \mathrm{y}(1) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$ in this proposition *implies* $\mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$ assumed by CFVAE. From the decomposition rule of conditional independence, we have $\mathrm{y}(0), \mathrm{y}(1) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} \implies \forall t (\mathrm{y}(t) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}) \implies \forall t, \bar{t} (\mathrm{y}(t) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} = \bar{t})$, 4 independence in total. Using only two of them, and with the consistency of counterfactual, we have $\forall t (\mathrm{y}(t) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} = t) \implies \forall t (\mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} = t) \implies \mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$. In general, the other two independence $\mathrm{y}(0) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} = 1$ and $\mathrm{y}(1) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} = 0$ do not hold, and thus $\mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t} \not\Longrightarrow \mathrm{y}(0), \mathrm{y}(1) \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$.

For any balancing covariates (including noiseless proxies), CFVAE can identify treatment effects, by learning a balancing score of the true confounder as latent representation.

Theorem 2 (Identifiability with balancing covariate)**.** *Assume $\mathbf{z}$ satisfies strong ignorability and let $\mathbf{x}$ be a balancing covariate of $\mathbf{z}$. Further specify the learning model* (7) *with $\boldsymbol{h}$, $\boldsymbol{k}$ depend on $\mathbf{x}$ but not* t *(i.e. $h_i(\boldsymbol{x}, t) = h_i(\boldsymbol{x})$ and same for $\boldsymbol{k}$). And in addition assume the same as Corollary 1, then, for* both $t = 0, 1$, *we have $\mathbf{z} = \mathrm{diag}(\boldsymbol{a})\mathbf{z}' + \boldsymbol{b} := \mathcal{A}(\mathbf{z}')$, and $\mathbf{z}'$ satisfies strong ignorability. We identify the potential outcomes by $\mu_{\bar{t}}(\boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\mathrm{y}|\mathbf{z}', \mathbf{x} = \boldsymbol{x}, \mathrm{t} = \bar{t})) = \mathbb{E}(\boldsymbol{f}'_{\bar{t}}(\boldsymbol{r}'_t(\boldsymbol{x}, \mathrm{y}_t))|\mathbf{x} = \boldsymbol{x})$.*

The result may be more easily understood as following. Now with the same affine transformation, for both treatment groups, given observation y_t, the counterfactual prediction given by CFVAE is the same as truth: $y'_{1-t} = \boldsymbol{f}'_{1-t}(\boldsymbol{f}'^{-1}_t(y_t)) = \boldsymbol{f}'_{1-t}(\boldsymbol{z}'_t) = \boldsymbol{f}_{1-t} \circ \mathcal{A}(\mathcal{A}^{-1}(\boldsymbol{z}_t)) = \boldsymbol{f}_{1-t}(\boldsymbol{z}_t) = y(1-t)$ (we use $\boldsymbol{f}'^{-1}_t(y_t) = \boldsymbol{r}'_t(\boldsymbol{x}, y_t) = \boldsymbol{z}'_t$ in second equality). We can identify potential outcomes by

$$\mu_{\bar{t}}(\boldsymbol{x}) = \mathbb{E}(\mathrm{y}(\bar{t})|\mathbf{x} = \boldsymbol{x}) = \mathbb{E}(\mathrm{y}'_{\bar{t}}|\mathbf{x} = \boldsymbol{x}) = \mathbb{E}(\boldsymbol{f}'_{\bar{t}}(\boldsymbol{r}'_t(\boldsymbol{x}, \mathrm{y}_t))|\mathbf{x} = \boldsymbol{x}) \tag{9}$$

Note that counterfactual assignment $\mathrm{t} = \bar{t}$ may or may not be the same as factual t. The algorithm for estimation CATE and ATE is as following. After training CFVAE, we feed data $\mathcal{D} = \{(\boldsymbol{x}, y_t) := (\boldsymbol{x}, y, t)\}$ into the encoder, and draw sample from it: $q(\mathbf{z}'|\mathbf{x} = \boldsymbol{x}, \mathrm{y} = y, \mathrm{t} = t) = \delta(\mathbf{z}' - \boldsymbol{r}'_t(\boldsymbol{x}, y_t))$. Then, setting $\mathrm{t} = \bar{t} \in \{0, 1\}$ in the decoder, feed posterior sample $\{\boldsymbol{z}'_t = \boldsymbol{r}'_t(\boldsymbol{x}, y_t)\}$, we get counterfactual prediction $p(\mathrm{y}'|\mathbf{z}' = \boldsymbol{z}'_t, \mathrm{t} = \bar{t}) = \delta(\mathrm{y}'_{\bar{t}} - \boldsymbol{f}'_{\bar{t}}(\boldsymbol{z}'_t))$. Finally, we estimate ATE by taking average $\mathbb{E}_{\mathcal{D}}(y'_1 - y'_0)$, and CATE by adding conditioning on $\boldsymbol{x}$ $\mathbb{E}_{\{\mathcal{D}|\mathbf{x}=\boldsymbol{x}\}}(y'_1 - y'_0)$.

A caveat is that (9) requires *post-treatment* observation y_t. Often, it is desirable that we can also have *pre-treatment* prediction for a new subject, with only the observation of its covariate $\mathbf{x} = \boldsymbol{x}$. To this end, we use conditional prior $p(\mathbf{z}'|\mathbf{x})$ as a pre-treatment predictor for $\mathbf{z}'$: input $\boldsymbol{x}$ and draw sample from $p(\mathbf{z}'|\mathbf{x} = \boldsymbol{x})$ instead of q, and all the others remain the same. Since the ELBO has a KL term between $p(\mathbf{z}'|\mathbf{x})$ and q, the two distributions should not be very different, and we will also have sensible pre-treatment estimation of treatment effects.

Although our method works under unobserved confounding, it still formally requires small outcome noise and balancing covariate. However, experiments show our method can work very well with large outcome noise, and the covariates can be non-balancing and also directly affect the outcome, including general proxies, IVs, and even networked data.

6 Experiments

As in previous works (Shalit et al., 2017; Louizos et al., 2017), we report the absolute error of ATE $\epsilon_{ATE} := |\mathbb{E}_{\mathcal{D}}(y(1) - y(0)) - \mathbb{E}_{\mathcal{D}}(y'_1 - y'_0)|$, and the square root of empirical PEHE (Hill, 2011) $\epsilon_{PEHE} := \mathbb{E}_{\mathcal{D}}((y(1) - y(0)) - (y'_1 - y'_0))^2$ for individual-level treatment effects.

Unless otherwise indicated, for each function $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{k}, \boldsymbol{r}, \boldsymbol{s}$ in (7), we use a multilayer perceptron (MLP) that has 3*200 hidden units with ReLU activation, and $\boldsymbol{h}, \boldsymbol{k}$ depend only on $\mathbf{x}$. The Adam optimizer with initial learning rate 10^{-4} and batch size 100 is employed. More details on hyper-parameters and experimental settings are given in each experiment, and are explained in Appendix.

All experiments use early-stopping of training by evaluating the ELBO on a validation set. We evaluate the post-treatment performance on training and validation set jointly (This is non-trivial. Recall the fundamental problem of causal inference). The treatment and (factual) outcome should not be observed for pre-treatment predictions, so we report them on a testing set.

6.1 Synthetic dataset

We generate data following (10) with z, y 1-dimensional and $\mathbf{x}$ 3-dimensional. μ_i and σ_i are randomly generated in range $(-0.2, 0.2)$ and $(0, 0.2)$, respectively. The functions h, k, l are linear with random coefficients. The outcome model is built for the two treatments separately, i.e. $f(\mathbf{z}, t) := f_t(\mathbf{z}), t = 0, 1$. We generate two kinds of outcome models, depending on the type of f_t: linear and nonlinear outcome models use random linear functions and NNs with random weights, respectively.

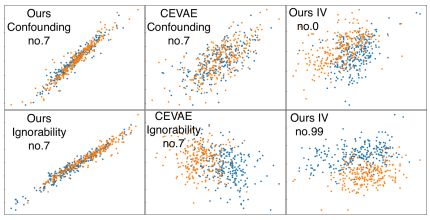

Figure 2: Plots of recovered (x) - true (y) latent on the nonlinear outcome. Blue: $t = 0$, Orange: $t = 1$. $\alpha, \beta = 0.4$. "no." indicates index among the 100 random models.

$$\begin{aligned} \mathbf{x} \sim \textstyle\prod_{i=1}^{3} \mathcal{N}(\mu_i, \sigma_i); \quad \mathrm{z}|\mathbf{x} \sim \mathcal{N}(h(\mathbf{x}), \beta k(\mathbf{x})); \\ \mathrm{t}|\mathbf{x}, \mathrm{z} \sim \mathrm{Bern}(\mathrm{Logistic}(l(\mathbf{x}, \mathrm{z}))); \quad \mathrm{y}|\mathrm{z}, \mathrm{t} \sim \mathcal{N}(C_{\mathrm{t}}^{-1} f(\mathrm{z}, \mathrm{t}), \alpha). \end{aligned} \tag{10}$$

We adjust the outcome and proxy noise level by α, β respectively. The output of f_t is normalized by $C_t := \mathrm{Var}_{\{\mathcal{D}|\mathrm{t}=t\}}(f_t(\mathrm{z}))$. This means we need to use $0 \leq \alpha < 1$ to have a reasonable level of noise on y (the scales of mean and variance are comparable). Similar reasoning applies to $\mathrm{z}|\mathbf{x}$; outputs of h, k have approximately the same range of values since the functions' coefficients are generated by the same weight initializer.

We experiment on three different causal settings (indicated in *Italic*). To introduce $\mathbf{x}$ as *IV*, we generate another 1-dimensional random source w in the same way as $\mathbf{x}$, and use w instead of $\mathbf{x}$ to generate $\mathrm{z}|\mathrm{w} \sim \mathcal{N}(h(\mathrm{w}), \beta k(\mathrm{w}))$. Besides taking inputs $\mathbf{x}, \mathrm{z}$ in l, we consider two special cases: $l := l(\mathbf{x})$ ($\mathbf{x}$ fully satisfies *ignorability*) and $l := l(\mathrm{z})$ (unobserved confounder z and *non-balancing proxy* $\mathbf{x}$ of z). Except indicated above, other aspects of the models are specified by (10).

In each causal setting, and with the same kind of outcome models and noise levels (α, β), we evaluate CFVAE and CEVAE on 100 random data generating models, with different sets of functions f, h, k, l in (10). For each model, we sample 1500 data points, and split them into 3 equal sets for training, validation, and testing. Both the methods use 1-dimensional latent variable in VAE. For fair comparison, all the hyper-parameters, including type and size of NNs, learning rate, and batch size, are the same for both the methods.

Figure 3 shows our method significantly outperforms CEVAE on all cases. Each method works the best under ignorability, as expected. The performances of our method on IV and proxy settings match that of CEVAE under ignorability, showing the effective deconfounding. Figure 2 shows our method learns highly interpretable representation as an approximate affine transformation of the true latent value. To our surprise, CEVAE is also possible to achieve this when both noises are small, though the quality of recovery is lower than CFVAE. The relationship to the true latent is significantly obscured under IVs, because the true latent is correlated to IVs only given t, while we model it by $p(\mathrm{z}'|\mathbf{x})$ as required by Theorem 2.

We can see our method and also CEVAE are very robust w.r.t. both outcome and proxy noise. This may due to the good probabilistic modeling of the noise by VAE. Still, we can see in Appendix that the noise level affects how well we recover the latent variable.

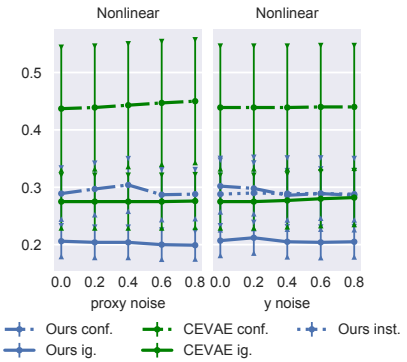

Figure 3: PEHE on nonlinear synthetic dataset. Error bar on 100 random models. We adjust one of the noise levels α, β in each panel, with another fixed to 0.2. See Appendix for results on linear outcome. Results for ATE and post-treatment are similar.

6.2 IHDP BENCHMARK DATASET

The IHDP dataset (Hill, 2011) is widely used to evaluate machine learning based causal inference methods, e.g. Shalit et al. (2017); Shi et al. (2019). Here, *ignorability* holds given the covariates. See Appendix for detailed descriptoins. Note, however, that this dataset violates our assumption $\mathrm{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$, since the covariates $\mathbf{x}$ directly affect the outcome. To overcome this, we add two components introduced by Shalit et al. (2017) into our method. First, we build two outcome functions $\boldsymbol{f}_t(\mathbf{z}), t = 0, 1$ in our learning model (7), using two separate NNs. Second, we add to our ELBO (6) a regularization term, which is the Wasserstein distance (Cuturi, 2013) between the learned $p(\mathbf{z}'|\mathbf{x}, \mathrm{t} = 0)$ and $p(\mathbf{z}'|\mathbf{x}, \mathrm{t} = 1)$. We find higher than 1-dimensional latent variable in CFVAE gives better results, very possibly due to the mismatched latent distribution: the confounder `race` is discrete but we use Gaussian latent variable. We report results with 10-dimensional latent variable.

As shown in Table 1, the proposed CFVAE matches the state-of-the-art methods under model misspecification. This robustness of VAE was also observed by Louizos et al. (2017), where they used 5-dimensional Gaussian latent variable to model a binary ground truth. And notably, without the two additional modifications, our method has the best ATE estimation and is overall the best among generative models (CEVAE, GANITE, and ours).

Table 1: Errors on IHDP. "A/B" means pre-treatment/post-treatment prediction. The mean and std are calculated over 1000 random draws of the data generating model. *Results with the two modifications. The results without the modifications are $\epsilon_{ATE} = \mathbf{.21}_{\pm.01}/\mathbf{.17}_{\pm.01}$ and $\sqrt{\epsilon_{PEHE}} = 1.0_{\pm.05}/.97_{\pm.04}$. **Bold** indicates method(s) that are *significantly* better than all the others. The results of the other methods are taken from Shalit et al. (2017), except GANITE, the result of which is from Yoon et al. (2018).

Method	**TMLE**	**BNN**	**CFR**	**CF**	**CEVAE**	**GANITE**	**Ours***
ϵ_{ATE}	NA/$.30_{\pm.01}$	$.42_{\pm.03}/.37_{\pm.03}$	$.27_{\pm.01}/.25_{\pm.01}$	$.40_{\pm.03}/18_{\pm.01}$	$.46_{\pm.02}/.34_{\pm.01}$	No data	$.31_{\pm.01}/.30_{\pm.01}$
$\sqrt{\epsilon_{PEHE}}$	NA/$5.0_{\pm.2}$	$2.1_{\pm.1}/2.2_{\pm.1}$	$\mathbf{.76}_{\pm.0}/\mathbf{.71}_{\pm.0}$	$.18_{\pm.01}/3.8_{\pm.2}$	$2.6_{\pm.1}/2.7_{\pm.1}$	$2.4_{\pm.4}/1.9_{\pm.4}$	$\mathbf{.77}_{\pm.02}/\mathbf{.69}_{\pm.02}$

6.3 POKEC SOCIAL NETWORK DATASET

Pokec (Leskovec & Krevl, 2014) is a real world social network dataset. We experiment on a semi-synthetic dataset based on Pokec, which was introduced in Veitch et al. (2019), and use exactly the same pre-processing and generating procedure. The pre-processed network has about 79,000 vertexes (users) connected by 1.3×10^6 undirected edges. The subset of users used here are restricted to three living districts that are within the same region. The network structure is expressed by binary adjacency matrix $\boldsymbol{G}$. Following Veitch et al. (2019), we split the users into 10 folds, test on each fold and report the mean and std of pre-treatment ATE predictions. We further separate the rest of users (in the other 9 folds) by $6:3$, for training and validation. Table 2 shows the results. Our method is

the best compared with the methods specialized for networked data. We report pre-treatment PEHE of our method in the Appendix, while Veitch et al. (2019) does not give individual-level prediction.

Table 2: Pre-treatment ATE on Pokec. Ground truth is 1. "Unadjusted" estimates ATE by $\mathbb{E}_{\mathcal{D}}(y_1) - \mathbb{E}_{\mathcal{D}}(y_0)$. "Parametric" is a stochastic block model for networked data (Gopalan & Blei, 2013). "Embed-" denotes the best alternatives given by Veitch et al. (2019). **Bold** indicates method(s) that are *significantly* better than all the others. 20-dimensional latent variable in CFVAE works better, and its result is reported. The results of the other methods are taken from Veitch et al. (2019).

	Unadjusted	Parametric	Embed-Reg.	Embed-IPW	Ours
`Age`	4.34 ± 0.05	4.06 ± 0.01	2.77 ± 0.35	3.12 ± 0.06	$\mathbf{2.08} \pm 0.32$
`District`	4.51 ± 0.05	3.22 ± 0.01	$\mathbf{1.75} \pm 0.20$	$\mathbf{1.66} \pm 0.07$	$\mathbf{1.68} \pm 0.10$
`Join Date`	4.03 ± 0.06	3.73 ± 0.01	2.41 ± 0.45	3.10 ± 0.07	$\mathbf{1.70} \pm 0.13$

Each user has 12 attributes, among which `district`, `age`, or `join date` is used as a confounder z to build 3 different datasets, with remaining 11 attributes used as covariate $\mathbf{x}$. Treatment t and outcome y are synthesised as following:

$$\mathrm{t} \sim \mathrm{Bern}(g(\mathrm{z})); \quad \mathrm{y} = \mathrm{t} + 10(g(\mathrm{z}) - 0.5) + \epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{11}$$

Note that `district` is of 3 categories; `age` and `join date` are also discretized into three bins. $g(\mathrm{z})$ maps these three categories and values to $\{0.15, 0.5, 0.85\}$.

Some assumptions to justify our method may not hold in this dataset. The important challenges are 1) $\mathbf{x}$ obviously does not satisfy ignorability, and 2) large outcome noise exists. On the other hand, given the huge network structure, most users can practically be identified by their attributes and neighborhood structure, which means z can be roughly seen as a deterministic function of $\boldsymbol{G}, \mathbf{x}$. Then, $\boldsymbol{G}, \mathbf{x}$ can be, as defined by us, *noiseless proxies* of z (see Appendix 8.5). CFVAE is then expected to control for the confounding to a large extent and able to learn a balancing score based on Theorem 2, if we can exploit the network structure effectively. This idea is comparable to Assumptions 2 and 4 in Veitch et al. (2019), which postulate directly that a balancing score can be learned in the limit of infinite large network.

To extract information from the network structure, we use Graph Convolutional Network (GCN) (Kipf & Welling, 2017) in conditional prior and encoder of CFVAE. A difficulty is that, the network $\boldsymbol{G}$ and covariates $\boldsymbol{X}$ of *all* users are always needed by GCN, regardless of whether it is in training, validation, or testing phase. However, the separation can still make sense if we take care that the treatment and outcome are used only in the respective phase, e.g., (y_m, t_m) of a testing user m is only used in testing. See Appendix for details.

7 Discussion

In this work, we proposed a new VAE architecture for estimating causal effects under unobserved confounding, with theoretical analysis and state-of-the-art performance. To the best of our knowledge, this is the first generative learning method that provably identifies treatment effects, without directly assuming that the true latent variable can be recovered. It is achieved by, on the one hand, noticing we only need the part of latent information that is correlated to treatment assignment, and, on the other hand, exploiting the recent advances that the latent variable can be recovered up to trivial transformations in a broad class of generative models.

Despite the formal requirement, the experiments show our method is robust to large outcome noise. Theoretical analysis of this phenomenon is an interesting direction for future work. A related theoretical issue is that, while Khemakhem et al. (2020) assumes fixed distribution of noise on $\mathbf{y}$, we observed that, in most cases, allowing the noise distribution to depend on $\mathbf{z}, \mathrm{t}$ improves performance. Extending identifiability to conditional noise models is also an interesting direction.

When the latent model is misspecified (Sec. 6.2 and 6.3), our method still matches the state-of-the-art, though we cannot see apparent relationship between recovered latent variable and the true one. It would be nice to see the learned representation indeed preserves causal properties under model misspecification, for example, by some causally-specialized metrics, e.g. Suter et al. (2019). Given the fact that all nonlinear ICA based identifiability requires an injective mapping between the latent and observed variables, theoretical extensions to discrete latent variable would be challenging.

REFERENCES

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Elizabeth S Allman, Catherine Matias, John A Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

Sander Greenland. The effect of misclassification in the presence of covariates. *American journal of epidemiology*, 112(4):564–569, 1980.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423, 2017.

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

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pp. 3765–3773, 2016.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868, 2019.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2281–2290, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection, 2014.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019*, pp. 45, 2019.

Elizabeth L Ogburn. Challenges to estimating contagion effects from observational data. In *Complex Spreading Phenomena in Social Systems*, pp. 47–64. Springer, 2018.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

Geoffrey Roeder, Luke Metz, and Diederik P Kingma. On linear identifiability of learned representations. *arXiv preprint arXiv:2007.00810*, 2020.

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

KJ Rothman, S Greenland, and TL Lash. Bias analysis. *Modern epidemiology*, 3, 2008.

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

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2507–2517, 2019.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pp. 3483–3491, 2015.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. In *Advances in Neural Information Processing Systems*, pp. 13792–13802, 2019.

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

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Pengzhou Wu and Kenji Fukumizu. Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. volume 108 of *Proceedings of Machine Learning Research*, pp. 1157–1167, Online, 26–28 Aug 2020. PMLR. URL `http://proceedings.mlr.press/v108/wu20b.html`.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2633–2643, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ByKWUeWA-`.

8 APPENDIX

8.1 PROOFS

Proof of Corollary 1. We need the *consistency*[3] of our VAE to learn a observational distribution equaling to the true one in the limit of infinite data, so that the learned parameters $\boldsymbol{\theta}'_t$ is in the equivalence class of $\boldsymbol{\theta}_t$ defined by (8). This can be proved (Khemakhem et al., 2020, Theorem 4) by assuming: 1) our VAE is flexible enough to ensure the ELBO is tight (equals to the log likelihood of our model) for some parameters; 2) the optimization algorithm can achieve the *global* maximum of ELBO (again equals to the log likelihood).

In this proof, all equations and variables should condition on t, and we omit the conditioning in notation for convenience. In the limit of $\boldsymbol{\sigma}_{\mathbf{y}} \to \mathbf{0}$, the decoder degenerates to a delta function: $p(\mathbf{y}|\mathbf{z}) = \delta(\mathbf{y} - \boldsymbol{f}(\mathbf{z}))$, we have $\mathbf{y} = \boldsymbol{f}(\mathbf{z})$ and $\mathbf{y}' = \boldsymbol{f}'(\mathbf{z}')$. From the consistency of VAE, $\mathbf{y}$ should have the same support as $\mathbf{y}'$. For all $\boldsymbol{y}$ in the support, there exist a *unique* $\boldsymbol{z}$ and a *unique* $\boldsymbol{z}'$ satisfy $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}'(\boldsymbol{z}')$ (use injectivity). Substitute $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z})$ into the l.h.s of (8), and $\boldsymbol{y} = \boldsymbol{f}'(\boldsymbol{z}')$ into the r.h.s, we have $\boldsymbol{z} = \mathrm{diag}(\boldsymbol{a})\boldsymbol{z}' + \boldsymbol{b}$. The relation is one-to-one for all $\boldsymbol{z}$, so we get $\mathbf{z} = \mathcal{A}(\mathbf{z}')$. Similar result for $\boldsymbol{f}$ follows. □

Proposition 2 (Properties of conditional independence)**.** *For random variables* $\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$*. We have (Pearl, 2009, 1.1.55):*

$$\begin{aligned} \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \wedge \mathbf{x} \perp\!\!\!\perp \mathbf{w}|\mathbf{y}, \mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{w}, \mathbf{y}|\mathbf{z} \textit{ (Contraction)}. \\ \mathbf{x} \perp\!\!\!\perp \mathbf{w}, \mathbf{y}|\mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{w}, \mathbf{z} \textit{ (Weak union)}. \\ \mathbf{x} \perp\!\!\!\perp \mathbf{w}, \mathbf{y}|\mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \textit{ (Decomposition)}. \end{aligned}$$

Proof of Proposition 1. This proof will use the above three properties of conditional independence. We first write our assumptions in conditional independence, as *A1.* $\mathrm{t} \perp\!\!\!\perp \mathbf{z}|\mathbf{x}$ (balancing covariate), *A2.* $\mathbf{w} \perp\!\!\!\perp \mathrm{t}|\mathbf{z}$ (ignorability given $\mathbf{z}$), and *A3.* $\mathbf{w} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}, \mathrm{t}$, where $\mathbf{w} := (\mathrm{y}(0), \mathrm{y}(1))$.

Now, from *A2* and *A3*, using contraction, we have $\mathbf{w} \perp\!\!\!\perp \mathbf{x}, \mathrm{t}|\mathbf{z}$, then using weak union, we have $\mathbf{w} \perp\!\!\!\perp \mathrm{t}|\mathbf{x}, \mathbf{z}$. From this last independence and *A1*, using contraction, we have $\mathrm{t} \perp\!\!\!\perp \mathbf{z}, \mathbf{w}|\mathbf{x}$. Then $\mathrm{t} \perp\!\!\!\perp \mathbf{w}|\mathbf{x}$ follows by decomposition. □

[3]This is the statistical consistency of an estimator. Do not confuse with the consistency of counterfactuals.

Proof of Theorem 2. From the proof of Theorem 1 in Khemakhem et al. (2020), we know $\boldsymbol{A}_t, \boldsymbol{b}_t$ depend on t only through $\boldsymbol{h}, \boldsymbol{k}$. But we assume $\boldsymbol{h}, \boldsymbol{k}$ do not depend on t. So we have $\mathbf{z} = \mathcal{A}(\mathbf{z}')$ for both $t = 0, 1$. From Theorem 1, $\mathbf{z}'$ is a balancing score of $\mathbf{z}$, and satisfies strong ignorability. Here, we proceed a bit different from (2), we have

$$\begin{aligned}\mu_{\bar{t}}(\boldsymbol{x}) &= \mathbb{E}(\mathbb{E}(\mathrm{y}(\bar{t})|\mathbf{z}', \mathbf{x} = \boldsymbol{x})) = \int \mathbb{E}(\mathrm{y}(\bar{t})|\mathbf{z}' = \boldsymbol{z}', \mathbf{x} = \boldsymbol{x}) p(\boldsymbol{z}'|\mathbf{x} = \boldsymbol{x}) d\boldsymbol{z}' \\ &= \int \mathbb{E}(\mathrm{y}|\mathbf{z}' = \boldsymbol{z}', \mathbf{x} = \boldsymbol{x}, \mathrm{t} = \bar{t}) p(\boldsymbol{z}'|\mathbf{x} = \boldsymbol{x}) d\boldsymbol{z}' = \int (\int p(y|\boldsymbol{z}', \boldsymbol{x}, \bar{t}) y dy) p(\boldsymbol{z}'|\boldsymbol{x}) d\boldsymbol{z}'\end{aligned} \tag{12}$$

Compare the rightmost side to (2), note that, there is no conditioning on t in $p(\boldsymbol{z}'|\boldsymbol{x})$, because we use the strong ignorability given $\mathbf{z}'$ (and consistency of counterfactuals) in the third equality, *after* expanding the outer expectation.

From the consistency of VAE, $p(\mathrm{y}|\boldsymbol{z}', \boldsymbol{x}, \bar{t}) = \delta(\mathrm{y} - \boldsymbol{f}'_{\bar{t}}(\boldsymbol{z}'))$, and $q(\mathbf{z}'|\boldsymbol{x}, y, t) = \delta(\mathbf{z}' - \boldsymbol{r}'_t(\boldsymbol{x}, y_t)) = p(\mathbf{z}'|\boldsymbol{x}, y, t) = \delta(\mathbf{z}' - \boldsymbol{f}'^{-1}_t(y_t))$ where $(\boldsymbol{x}, y_t) := (\boldsymbol{x}, y, t)$ is a data point. And $p(\boldsymbol{z}'|\boldsymbol{x}) = \sum_t \int_y p(\mathbf{z}'|\boldsymbol{x}, y, t) p(y, t|\boldsymbol{x}) dy = \sum_t \int_y \delta(\mathbf{z}' - \boldsymbol{f}'^{-1}_t(y_t)) p(y, t|\boldsymbol{x}) dy$. We have

$$\begin{aligned}\mu_{\bar{t}}(\boldsymbol{x}) &= \int_{z'} (\boldsymbol{f}'_{\bar{t}}(\boldsymbol{z}') \sum_t \int_y \delta(\mathbf{z}' - \boldsymbol{f}'^{-1}_t(y_t)) p(y, t|\boldsymbol{x}) dy) d\boldsymbol{z}' \\ &= \sum_t \int_y \boldsymbol{f}'_{\bar{t}}(\boldsymbol{f}'^{-1}_t(y_t)) p(y, t|\boldsymbol{x}) dy = \mathbb{E}(\boldsymbol{f}'_{\bar{t}}(\boldsymbol{r}'_t(\boldsymbol{x}, \mathrm{y}_t))|\mathbf{x} = \boldsymbol{x})\end{aligned} \tag{13}$$

We should note that $p(\mathbf{z}'|\boldsymbol{x}, y, t) := p_{\boldsymbol{\theta}'_t}(\mathbf{z}'|\mathbf{x} = \boldsymbol{x}, \mathrm{y} = y, \mathrm{t} = t) = p_{\boldsymbol{\theta}'_t}(\mathrm{y} = y, \mathbf{z}'|\mathbf{x} = \boldsymbol{x}, \mathrm{t} = t) / \int p_{\boldsymbol{\theta}'_t}(\mathrm{y} = y, \boldsymbol{z}'|\mathbf{x} = \boldsymbol{x}, \mathrm{t} = t) d\boldsymbol{z}'$ might *not* be equal to the truth $p(\mathbf{z}|\boldsymbol{x}, y, t) := p_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x} = \boldsymbol{x}, \mathrm{y} = y, \mathrm{t} = t)$ (in particular it is possible that $\boldsymbol{f}'_t \neq \boldsymbol{f}_t$), but they are in the same equivalence class in the sense that $\boldsymbol{\theta}'_t, \boldsymbol{\theta}_t$ should satisfy (8).

Also note that the learning of inverse mapping $\boldsymbol{f}'^{-1}_t$ in the encoder q is *enforced* by consistency ($q(\mathbf{z}'|\boldsymbol{x}, y, t) = p(\mathbf{z}'|\boldsymbol{x}, y, t)$), we can just use an MLP for $\boldsymbol{r}'$ in the encoder, and we will have $\boldsymbol{r}'_t = \boldsymbol{f}'^{-1}_t$ if the MLP is flexible enough to contain $\boldsymbol{f}'^{-1}_t$. Similar situations prevail when identifiability is achieved by nonlinear ICA (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020). □

8.2 On the three identification conditions

Ignorability given $\mathbf{z}$ means there is no correlation between factual assignment of treatment and counterfactual outcomes given $\mathbf{z}$, just as it is the case in RCT. Thus, it can be understood as *unconfoundedness* given $\mathbf{z}$, and $\mathbf{z}$ can be seen as the *confounder(s)* we want to control for. Positivity says the supports of $p(\mathrm{t} = t|\mathbf{x} = \boldsymbol{x}), t = 0, 1$ should be *overlapped*, and this ensures there are no impossible events in the conditions after adding $\mathrm{t} = t$, and the expectations can thus be estimated from observational data. Finally, consistent counterfactuals are well defined: given assignment of treatment $\mathrm{t} = t$, the *observational outcome* y should take the same value as the potential outcome $\mathrm{y}(t)$.

8.3 Conditional VAE

By adding a conditioning variable $\mathbf{c}$ (usually a class label), Conditional VAE (CVAE) (Sohn et al., 2015; Kingma et al., 2014) can give better reconstruction of observation of each class. The variational lower bound is

$$\log p(\mathbf{y}|\mathbf{c}) \geq \mathbb{E}_{\boldsymbol{z} \sim q} \log p(\mathbf{y}|\boldsymbol{z}, \mathbf{c}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}, \mathbf{c}) \| p(\mathbf{z}|\mathbf{c})) := \mathcal{L}_{CVAE}(\mathbf{y}, \mathbf{c}) \tag{14}$$

The conditioning on $\mathbf{c}$ in the prior is usually omitted, since the dependence between $\mathbf{c}$ and the latent representation is also involved in the encoder q.

8.4 Identifiability of representation in sec. 5.1 is not enough

Consider how the *recovered* $\mathbf{z}'$ would be used. For a control group ($t = 0$) data point $(\boldsymbol{x}, y, 0)$, the real challenge is to predict the counterfactual outcome $y(1)$. Taking the observation, the encoder will

output a posterior sample point $\boldsymbol{z}_0' = \boldsymbol{f}_0'^{-1}(y) = \mathcal{A}_0^{-1}(\boldsymbol{z}_0)$ (with zero outcome noise, the encoder degenerates to a delta function: $q(\mathbf{z}|\boldsymbol{x}, y, 0) = \delta(\mathbf{z} - \boldsymbol{f}_0'^{-1}(y))$). Then, we should do *counterfactual inference*, using decoder with counterfactual assignment $t = 1$: $y_1' = \boldsymbol{f}_1'(\boldsymbol{z}_0') = \boldsymbol{f}_1 \circ \mathcal{A}_1(\mathcal{A}_0^{-1}(\boldsymbol{z}_0))$. This prediction can be arbitrary far from the truth $y(1) = \boldsymbol{f}_1(\boldsymbol{z}_0)$, due to the difference between $\mathcal{A}_1$ and $\mathcal{A}_0$. More concretely, this is because when learning the decoder, only the posterior sample of the treatment group ($t = 1$) is fed to $\boldsymbol{f}_1'$, and the posterior sample is different to the true value by the affine transformation $\mathcal{A}_1$, while it is $\mathcal{A}_0$ for $\boldsymbol{z}_0'$.

8.5 Two special case of balancing covariate

Definition 2 (Noiseless proxy)**.** Random variable $\mathbf{x}$ is a noiseless proxy of random variable $\mathbf{z}$ if $\mathbf{z}$ is a function of $\mathbf{x}$ ($\mathbf{z} = \boldsymbol{\omega}(\mathbf{x})$).

Noiseless proxy is a special case of balancing covariate because if $\mathbf{x} = \boldsymbol{x}$ is given, we know $\boldsymbol{z} = \boldsymbol{\omega}(\boldsymbol{x})$ and $\boldsymbol{\omega}$ is a deterministic function, then $p(\mathbf{z}|\mathbf{x} = \boldsymbol{x}) = p(\mathbf{z}|\mathbf{x} = \boldsymbol{x}, \mathrm{t}) = \delta(\mathbf{z} - \boldsymbol{\omega}(\boldsymbol{x}))$. Also note that, a noiseless proxy always has higher dimensionality than $\mathbf{z}$, or at least the same.

Intuitively, if the value of $\mathbf{x}$ is given, there is no further uncertainty about $\boldsymbol{z}$, so the observation of $\boldsymbol{x}$ may work equally well to adjust for confounding. But, as we will see soon, a noiseless proxy of the true confounder does *not* satisfy positivity.

Definition 3 (Injective proxy)**.** Random variable $\mathbf{x}$ is an injective proxy of random variable $\mathbf{z}$ if $\mathbf{x}$ is an injective function of $\mathbf{z}$ ($\mathbf{x} = \boldsymbol{\chi}(\mathbf{z})$, $\boldsymbol{\chi}$ is injective).

Injective proxy is again a special case of noiseless proxy, since, by injectivity, $\mathbf{z} = \boldsymbol{\chi}^{-1}(\mathbf{x})$, i.e. $\mathbf{z}$ is also a function of $\mathbf{x}$.

Under this very special case, that is, if $\mathbf{x}$ is an injective proxy of the true confounder $\mathbf{z}$, we finally have $\mathbf{x}$ is a balancing score and satisfies strong ignorability, since $\mathbf{x}$ is a balancing covariate and a function of $\mathbf{z}$. To see this in another way, let $f = e \circ \boldsymbol{\chi}^{-1}$ and $\boldsymbol{b} = \boldsymbol{\chi}$ in Theorem 1, then $f(\mathbf{x}) = f(\boldsymbol{b}(\mathbf{z})) = e(\mathbf{z})$. By strong ignorability of $\mathbf{x}$, (2) has a simpler counterpart $\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathrm{y}(t)|\mathbf{x} = \boldsymbol{x}) = \mathbb{E}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}, \mathrm{t} = t)$. Thus, a regression of y on $(\mathbf{x}, \mathrm{t})$ will give a valid estimator of CATE and ATE.

However, a noiseless but *non*-injective proxy is *not* a balancing score, in particular, positivity might *not* hold. Here, a simple regression will not do. This is exactly because $\boldsymbol{\omega}$ is non-injective, hence multiple values of $\mathbf{x}$ that cause non-overlapped supports of $p(\mathrm{t} = t|\mathbf{x} = \boldsymbol{x}), t = 0, 1$ might be mapped to the same value of $\mathbf{z}$. An extreme example would be $\mathrm{t} = \mathbb{I}(\mathrm{x} > 0), \mathrm{z} = |\mathrm{x}|$. We can see $p(\mathrm{t} = t|\mathrm{x})$ are totally non-overlapped, but $\forall t, z \neq 0 : p(\mathrm{t} = t|\mathrm{z} = z) = 1/2$.

8.6 Details and additional results for experiments

8.6.1 Synthetic data

Interestingly, linear outcome models seem harder for both methods, maybe because the two true linear outcome models for $t = 0, 1$ are more similar, and it is harder to distinguish and learning outcome models. Note that after generating the outcomes and before the data is used, we normalize the distribution of ATE of the 100 generating models, so the errors on linear and nonlinear settings are basically comparable.

You can find more plots for latent recovery at the end of the paper.

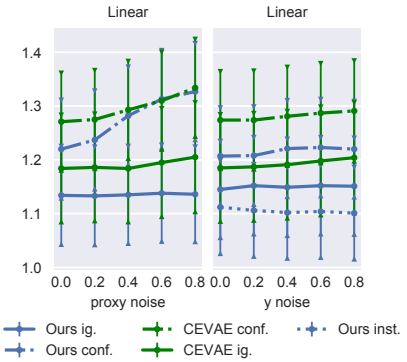

Figure 4: PEHE on linear synthetic dataset. Error bar on 100 random models. We adjust one of α, β at a time. Results for ATE and post-treatment are similar.

8.6.2 IHDP

IHDP is based on an RCT where each data point represents a child with 25 features about their birth and mothers. Race is introduced as a confounder by artificially removing all treated children with nonwhite

mothers. There are 747 subjects left in the dataset. The outcome is synthesized by taking the covariates (features excluding race) as input, hence *ignorability* holds given the covariates. Following previous work, we split the dataset by 63:27:10 for training, validation, and testing.

8.6.3 POKEC

GCN takes the network matrix $\boldsymbol{G}$ and the *whole* covariates matrix $\boldsymbol{X} := (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_M^T)^T$, where M is user number, and outputs a representation matrix $\boldsymbol{R}$, again for all users. During training, we *select* the rows in $\boldsymbol{R}$ that correspond to users in training set. Then, treat this *training representation matrix* as if it is the covariates matrix for a non-networked dataset, that is, the downstream networks in conditional prior and encoder are the same as in the above two experiments, but take $(\boldsymbol{R}_{m,:})^T$ where $\boldsymbol{x}_m$ was expected as input. And we have respective selection operations for validation and testing. We can still train CFVAE including GCN by Adam, simply setting the gradients of non-seleted rows of $\boldsymbol{R}$ to 0.

Note that GCN cannot be trained using mini-batch, instead, we perform batch gradient decent using full dataset for each iteration, with initial learning rate 10^{-2}. We use dropout (Srivastava et al., 2014) with rate 0.1 to prevent overfitting.

The pre-treatment $\sqrt{\epsilon_{PEHE}}$ for `Age`, `District`, and `Join date` confounders are 1.085, 0.686, and 0.699 respectively, practically the same as ATE error.

8.7 ADDITIONAL PLOTS ON SYNTHETIC DATASETS

See next pages.

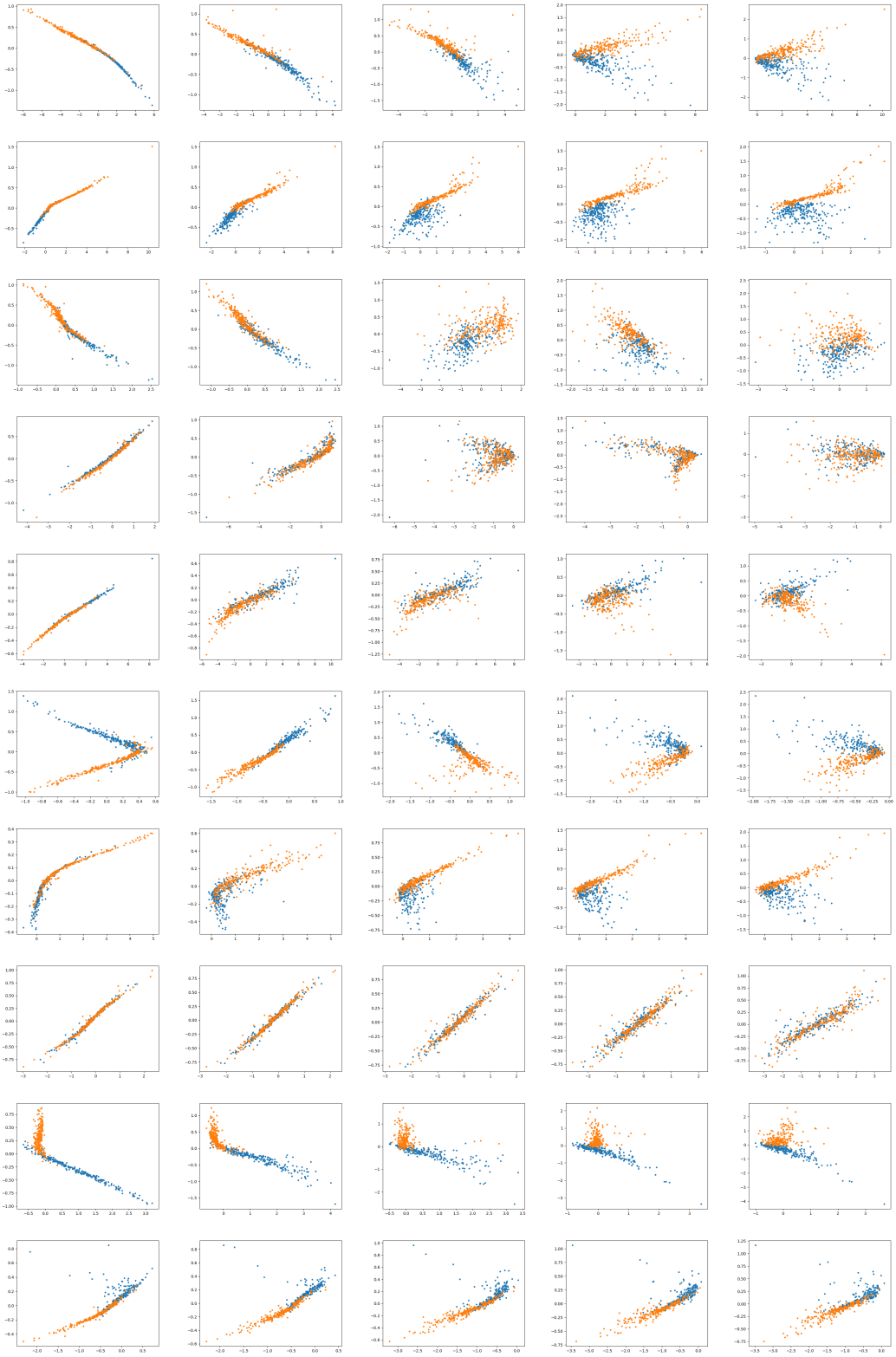

Figure 5: Plots of recovered-true latent under *unobserved confounding*. Rows: first 10 nonlinear random models, columns: *proxy* noise level.

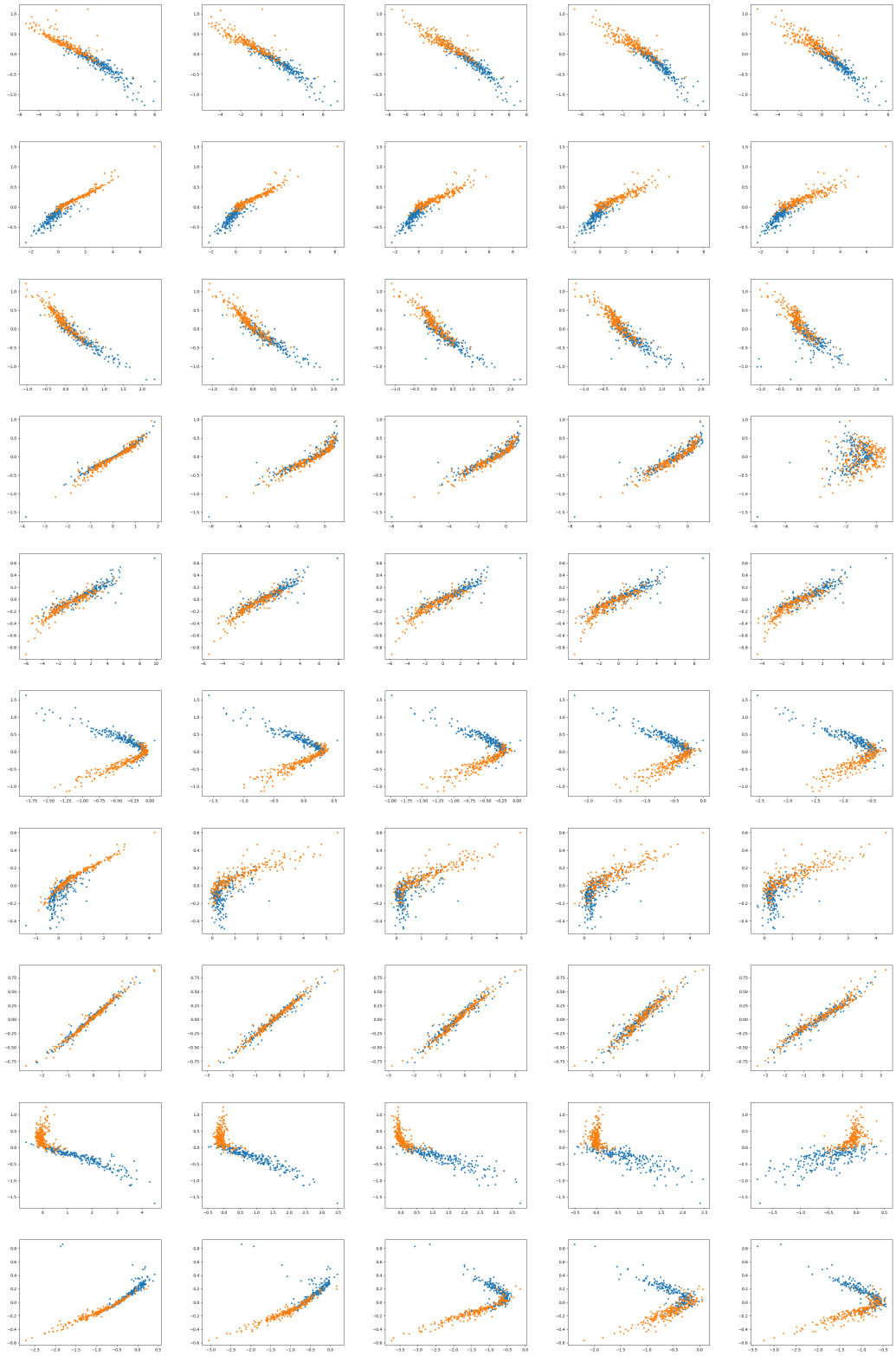

Figure 6: Plots of recovered-true latent under *unobserved confounding*. Rows: first 10 nonlinear random models, columns: *outcome* noise level.

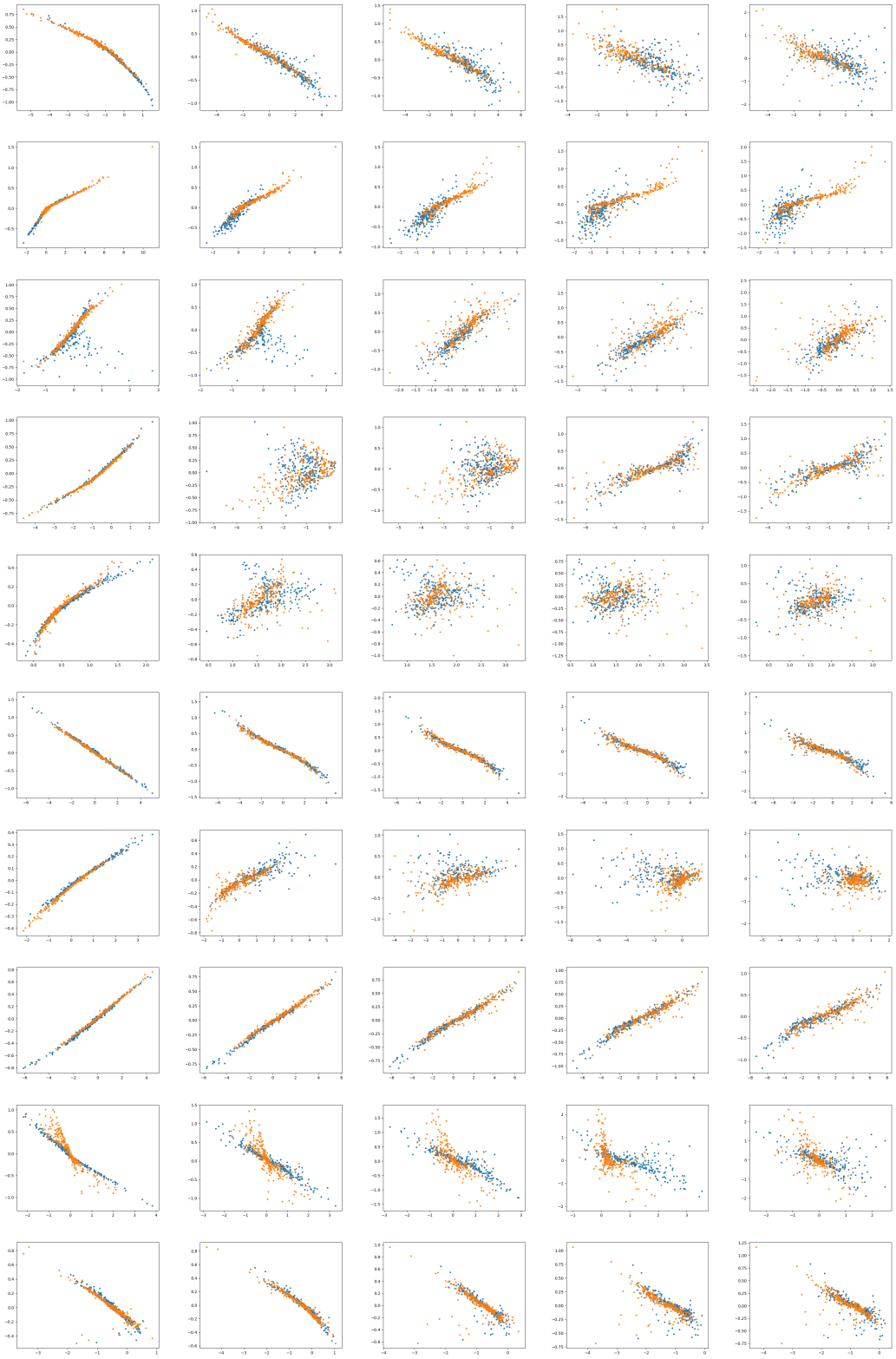

Figure 7: Plots of recovered-true latent when *ignorability* holds. Rows: first 10 nonlinear random models, columns: *proxy* noise level.

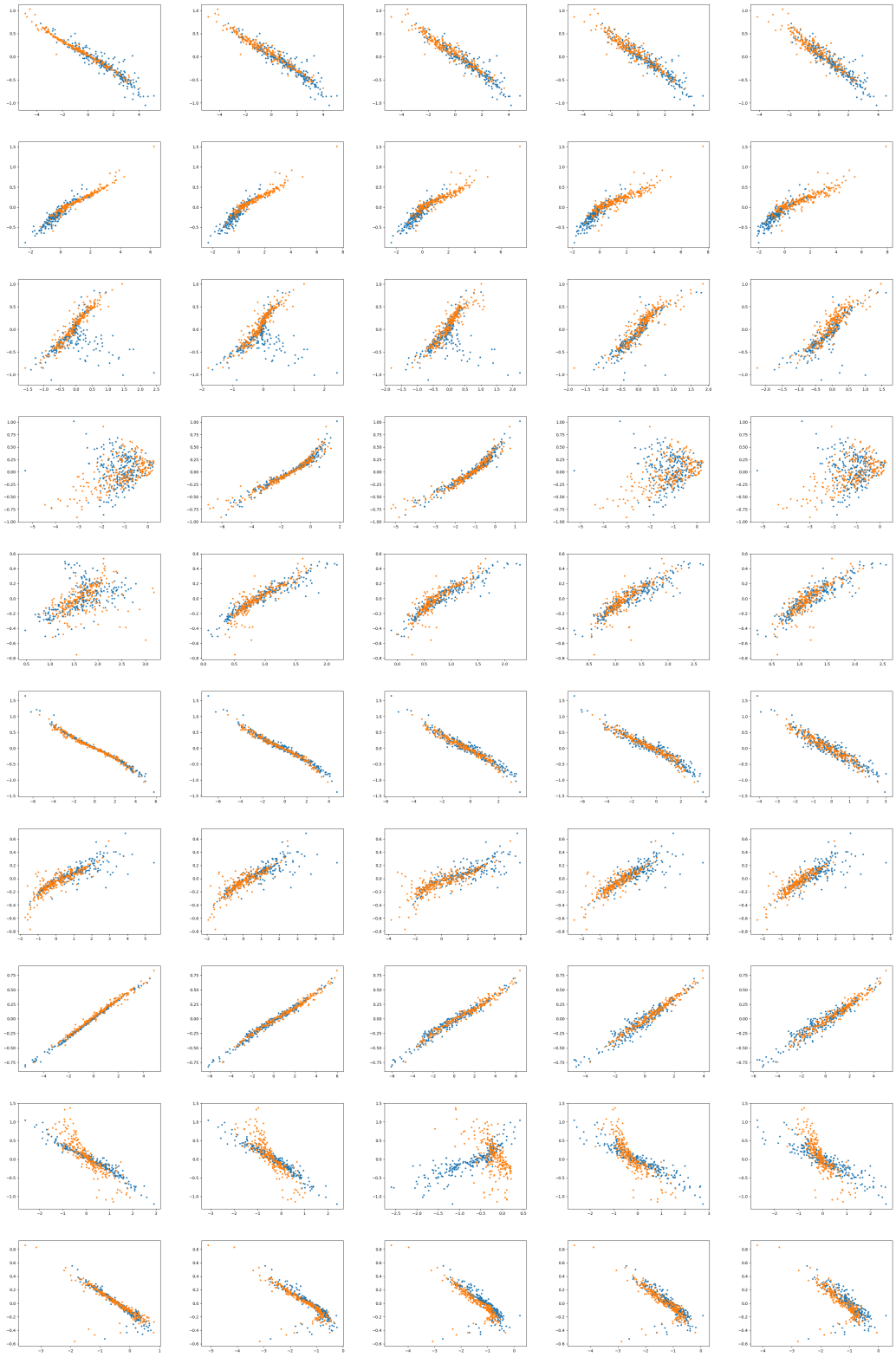

Figure 8: Plots of recovered-true latent when *ignorability* holds. Rows: first 10 nonlinear random models, columns: *outcome* noise level.

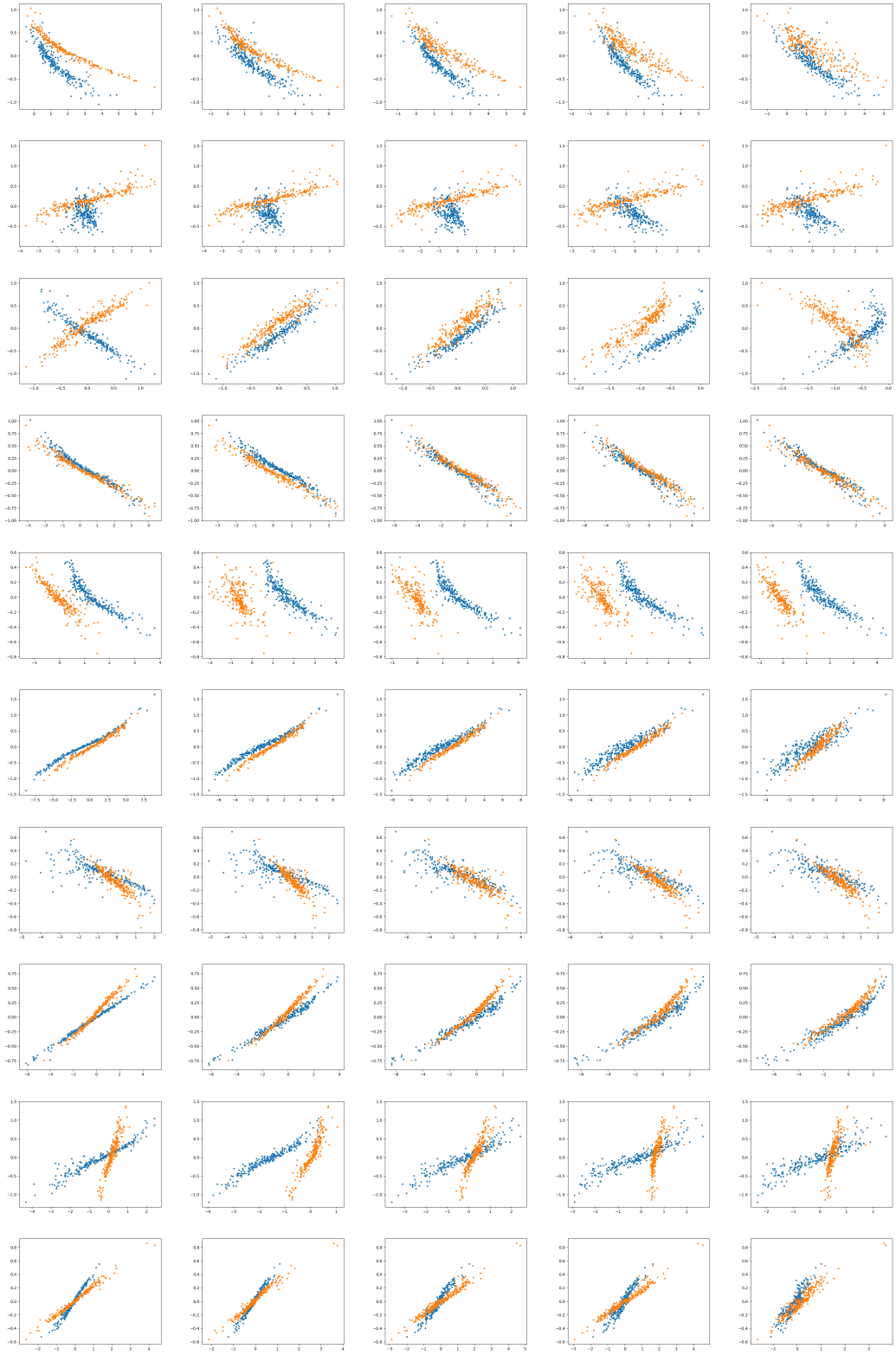

Figure 9: Plots of recovered-true latent when *ignorability* holds. Conditional prior *depends* on t. Rows: first 10 nonlinear random models, columns: *outcome* noise level. Compare to the previous figure, we can see the transformations for $t = 0, 1$ are *not* the same, confirming our Theorem 2.

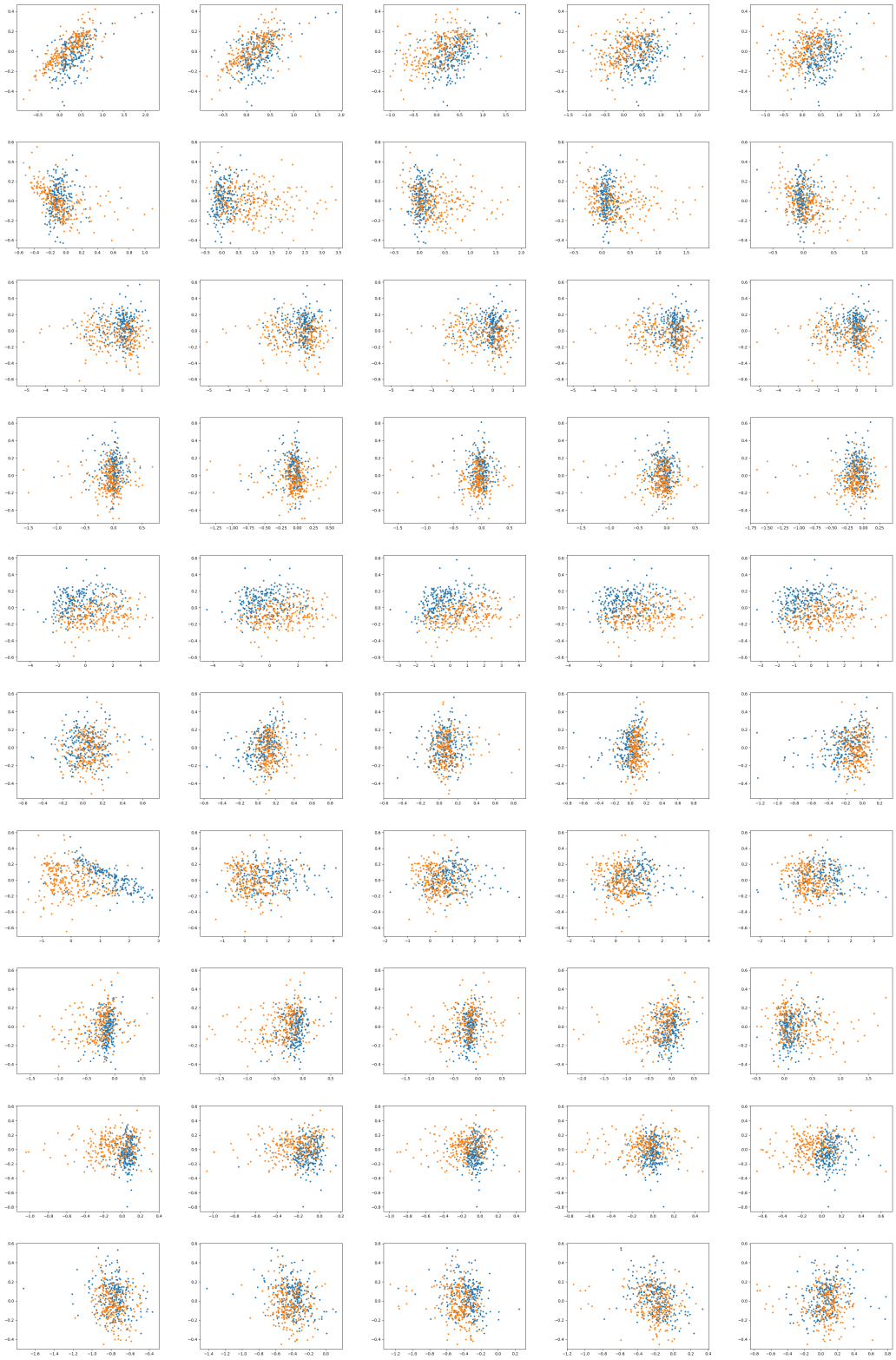

Figure 10: Plots of recovered-true latent on *IVs*. Rows: first 10 nonlinear random models, columns: *outcome* noise level.