Factual Observation Based Heterogeneity Learning for Counterfactual Prediction

Hao Zou ZOUH18@MAILS.TSINGHUA.EDU.CN
Department of Computer Science and Technology, Tsinghua University, Beijing, China

Haotian Wang ACCWHT@HOTMAIL.COM
National University of Defense Technology, Changsha, China

Renzhe Xu XRZ199721GMAIL.COM
Department of Computer Science and Technology, Tsinghua University, Beijing, China

Bo Li LIBO@SEM.TSINGHUA.EDU.CN
School of Economics and Management, Tsinghua University, Beijing, China

Jian Pei J.PEI@DUKE.EDU
Duke University, USA

Junjian Ye YEJUNJIAN@HUAWEI.COM
Huawei Noah's Ark Lab, Shenzhen, China

Peng Cui CUIP@TSINGHUA.EDU.CN
Department of Computer Science and Technology, Tsinghua University, Beijing, China

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Extant causal methods exclusively exploit the heterogeneity based on the observed covariates for heterogeneous outcome prediction. Even with nowadays big data, the collected covariates may not contain complete confounders. When some confounders are absent, the methods can suffer from confounding bias and missing heterogeneity. To address these two issues, we propose to leverage the factual observation in the observational data to recover the latent confounders. Since the learned confounder representation exploits the heterogeneity of latent confounders, it leads to finer granular heterogeneous outcome prediction, which is closer to the individual-level than prediction conditional on only covariates. Specifically, we propose a novel Factual Observation based Heterogeneity Learning (FOHL) algorithm with an encoder for confounder representation learning and a decoder for outcome prediction. Theoretical analysis reveals the validity of recovering confounders from factual observations to make the heterogeneous prediction closer to the individual-level. Furthermore, experimental results demonstrate that our FOHL method can outperform the existing baselines.

Keywords: Latent Confounder Recovery; Missing Heterogeneity; Closer to Individual-level

1. Introduction

Predicting the counterfactual outcomes of different treatments is of significant importance in many applications relevant to decision-making (Bica et al., 2020a; Bottou et al., 2013; Li et al., 2015). The common practice is to train an outcome predictive model by leveraging the observational datasets which are common and cheap (Hassanpour and Greiner, 2020, 2019; Johansson et al., 2016; Shalit et al., 2017; Yao et al., 2018). It has been supported by previous literature that the counterfactual outcome of treatments varies in different parts of the population and identifying heterogeneous outcomes can help improve the effectiveness of decisions (Lee et al., 2020). Therefore, it is of urgent

© 2023 H. Zou, H. Wang, R. Xu, B. Li, J. Pei, J. Ye & P. Cui.

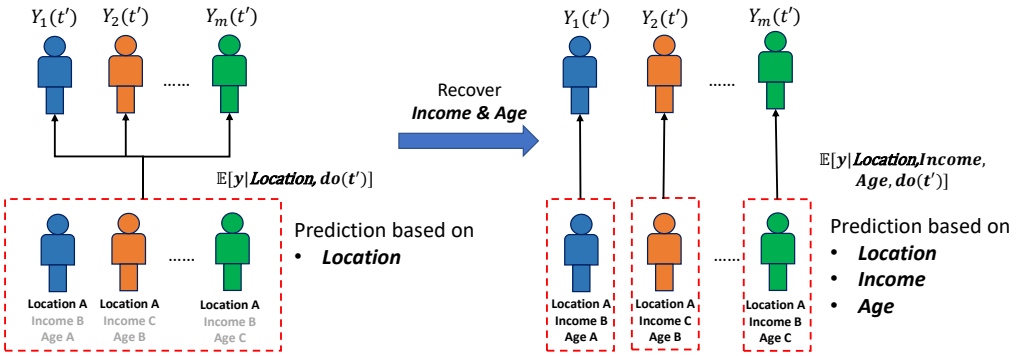


Figure 1: The diagram of heterogeneous outcome estimation conditional on observed covariates (i.e. location) v.s. complete confounders (i.e. all the three factors). When conditioning on only location, the resulting estimation (the average of m individual outcome) ignores the outcome variation among the sub-population. When we recover the latent confounders (i.e. income and age), we can obtain finer granular heterogeneous outcome prediction which steps closer to the individual-level.

need in many applications to estimate the heterogeneous response of treatments, such as precision medicine (Dahabreh et al., 2016) and personalized recommendations (He et al., 2017; Rendle, 2010). To resolve this task, some papers estimate the heterogeneous counterfactual outcome by partitioning the population based on the observed covariates which characterize the individuals to some extent (i.e. expected outcome conditional on observed covariates) (Shalit et al., 2017; Johansson et al., 2016; Yao et al., 2018; Zou et al., 2020; Hassanpour and Greiner, 2020, 2019; Assaad et al., 2021; Qian et al., 2021; Bica et al., 2020b; Yoon et al., 2018).

However, due to the limitation in information acquisition, some confounder variables which affect both outcome and treatments practically are absent in the observed covariates $\mathbf{X}$. This phenomenon of missing confounders brings two challenges to the predictive methods: confounding bias and **missing heterogeneity** which is a new perspective provided by us and also the focus of this paper. The former challenge, namely the confounding bias, has been investigated by many methods using different tools such as instrumental variables and negative controls (Hartford et al., 2017; Heckman, 1997; Miao et al., 2018; Miao and Tchetgen Tchetgen, 2017). However, the methods mentioned above only estimate the expectation of outcome (conditional on the observed covariates) formally $\mathbb{E}[\mathbf{y}|\mathbf{X}, do(\mathbf{t})]$, which is an incomplete characterization of one individual. As the missing confounders also play a vital role to depict the individualized causal effect, we define the missing heterogeneity problem as the variation of outcome in the sub-population with the same covariates $\mathbf{X}$ due to the heterogeneity of the missing confounders. Therefore, simply using the expected outcome value of the sub-population $\mathbb{E}[\mathbf{y}|\mathbf{X}, do(\mathbf{t})]$ as the estimation result and ignoring the variation of outcomes between individuals can lead to extra predictive error for individuals. To address the confounding bias and further the missing heterogeneity problem, it is a sensible idea to recover the complete confounders $\mathbf{Z}$ and then estimates heterogeneous outcome conditional on them $\mathbb{E}[\mathbf{y}|\mathbf{Z}, do(\mathbf{t})]$ which is at a finer granular level

and **steps closer to individual-level counterfactual prediction**. One example of the comparison of the heterogeneous outcomes at different granular levels is shown in Figure 1.

Fortunately, since the confounders affect the generation of factual treatments and outcomes in the observational dataset, some information about the latent confounders is encoded into them. Therefore, the factual observations (i.e. previously assigned treatments and particular outcomes) in the dataset present an opportunity to recover the latent confounders for simultaneously overcoming the confounding bias and missing heterogeneity problem. Drawing support from the rapidly developed latent variable models, we can model the underlying distribution of treatments, covariates, and outcomes with confounders, and then learn the representation of latent confounders from them. By conditioning on the learned confounder representation, we are capable of alleviating the missing heterogeneity problem and achieving finer granular heterogeneous outcome prediction than prediction conditional on only covariates. Notably, our idea of latent representation learning is inspired by the abduction step in solving the counterfactual query of the retrospective hypothetical scenarios in causal ladders (Pearl, 2009b, 2019). Although some works (Pawlowski et al., 2020; Khemakhem et al., 2021) have investigated this counterfactual query by inferring the latent exogenous noise in the framework of structure causal models (SCM) (Pearl, 2009a), they neglect the confounding bias problem and are restricted to some strong assumptions on SCM. For example, they need to access the full observations on the ancestor variables (corresponding to confounders and treatments in our problem).

In this paper, we consider the setting of multiple treatments, which is ubiquitous in reality (Qian et al., 2021; Zou et al., 2020; Wang and Blei, 2019). To simultaneously overcome the confounding bias and missing heterogeneity, we develop a novel Factual Observation based Heterogeneity Learning (FOHL) algorithm for recovering the latent confounders. The model is built upon the technique of variational inference (Khemakhem et al., 2020; Kingma and Welling, 2013; Rezende et al., 2014) to model the underlying distribution of the observed covariates, latent confounders, treatments and outcomes. With the encoder component, we can infer the latent confounder representation from factual observations. With the decoder component, we can predict heterogeneous outcomes at a finer granular heterogeneity level than prediction conditional on covariates by feeding into the inferred confounder representation and counterfactual treatments. Creatively, we set the covariates as the ancestor of latent confounders to practically avoid the complicated assumption on the covariate distribution and resist the influence of observational noise. Theoretical analysis shows the validity of our strategy. We also empirically conduct extensive experiments on the synthetic datasets and semi-synthetic datasets to show that our method outperforms the existing baselines.

The main contribution of this paper are summarized as follows:

- To the best of our knowledge, we are the first to investigate the missing heterogeneity problem under the setting of missing confounders. We propose a novel and easy-to-implement Factual Observation based Heterogeneity Learning (FOHL) algorithm to overcome both this problem and confounding bias by learning latent confounder representation.

- Theoretical analysis reveals that our strategy can remove confounding bias and more importantly alleviate missing heterogeneity for finer granular heterogeneous prediction than prediction conditional on only covariates. We conduct extensive experiments on both synthetic datasets and semi-synthetic datasets to show the effectiveness of our FOHL method.

2. Related Works

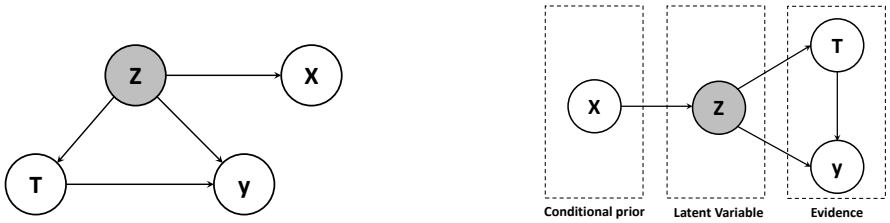
2.1. Heterogeneous outcome prediction conditional on covariates

There have been a large number of works that predict heterogeneous treatment outcomes conditional on covariates. One important branch of literature considers the setting without missing confounders. The effort of these methods is mainly devoted to reducing the correlation between the treatments and confounders. Motivated by the ideas in domain adaptation (Tzeng et al., 2014; Ganin and Lempitsky, 2015; Bousmalis et al., 2016), some methods propose the paradigm of learning treatment invariant representation to remove the correlation between the treatments and confounders and predict the outcome based on the learned representation (Shalit et al., 2017; Johansson et al., 2016; Yao et al., 2018; Bica et al., 2020a; Xu et al., 2021; Berrevoets et al., 2020; Zeng et al., 2020). Since over-enforcing the balancing property of representation may harm the predictive power (Assaad et al., 2021), sample re-weighting is an alternative solution (Zou et al., 2020; Hassanpour and Greiner, 2020, 2019; Lim, 2018; Assaad et al., 2021) to make the treatments and confounders independent in the re-weighted dataset. In addition, some papers resort to data augmentation to generate the counterfactual data points (Qian et al., 2021; Bica et al., 2020b; Yoon et al., 2018).

When faced with missing confounders, many methods are proposed to overcome the confounding bias and estimate the treatment outcome (conditional on covariates). Instrumental variables (Hartford et al., 2017; Heckman, 1997) and negative controls (Miao and Tchetgen Tchetgen, 2017; Miao et al., 2018) are two classical tools to remove the impact of missing confounders. However, these variables with restricted assumptions that are hard to seek in practice. In contrast, Louizos et al. (2017) proposes to infer the missing confounders from the proxies. Under the setting of multiple treatments, Wang and Blei (2019) observe that dependencies among the treatments can be leveraged to infer the missing confounders for removing confounding bias and obtaining unbiased treatment effect estimation. Miao et al. (2022) also gives two strategies and proves the theoretical validity with some additional assumptions. Nevertheless, these works still ignore the missing heterogeneity problem and integrate the effect of latent confounders in average/conditional treatment effect estimation.

2.2. Counterfactual Inference with Exogenous Noise Abduction

Answering counterfactual queries belongs to the third ladder of causality (Pearl, 2019), which is the most complicated and difficult one. This problem is usually resolved in the framework of the structural causal model (SCM) with three steps: abduction, action, and prediction (Pearl, 2009b). With the step of exogenous noise abduction, we can achieve a more accurate prediction for each instance than the intervention query (i.e. the second ladder of causality). Owing to the rapid development of deep generative models (Kingma and Welling, 2013; Rezende and Mohamed, 2015; Goodfellow et al., 2020), many methods utilize these flexible models to build the relationships between the variables and exogenous noises with weak specifications. Pawlowski et al. (2020) proposes three mechanisms to specify the structure equations and utilizes normalization flows, variational auto-encoders (VAE), and generative adversarial networks (GAN) to resolve them respectively. Based on the significant progress in generative energy-based models recently (Song et al., 2021; Ho et al., 2020), Sanchez and Tsaftaris (2021) propose to use the advanced diffusion model to construct the deep structural causal model. Khemakhem et al. (2021) generalizes the classical additive noise models (Hoyer et al., 2008) and proposes to use autoregressive flow models (Huang et al., 2018) to specify the equation which guarantees causal identifiability under some additional assumptions.



(*a*) Data generation process

(*b*) Probabilistic graph of our model

Figure 2: The graph of (a) the data generation process and (b) the probabilistic graph of our model. The grey circle means latent variables, and which white circle means observed variables.

The methodologies above assume that the exogenous noise is independent of the endogenous variables. However, this assumption is usually overly restrictive when our heterogeneous outcome prediction problem is formulated into this framework. The latent confounders in our problem is likely to be correlated with the observed covariates as well as treatments which further results in confounding bias. Therefore, it is inappropriate to simply formulate the latent confounders as the exogenous noise and solve the problem in the SCM framework.

3. Problem Formulation

In this paper, we investigate the heterogeneous outcome prediction for individuals based on observational datasets. The observational datasets consists of the treatments variables $\mathbf{T} \in \mathcal{T} = \{0,1\}^d$, outcomes $\mathbf{y} \in \mathbb{R}$, and the observed covariates $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ which can contains some confounder information. We mainly consider that setting that the covariates $\mathbf{X}$ are the noisy observation of latent confounders in this paper. For example, researchers usually can not obtain the true mental state of people but can observe it through questionnaire. For the brevity of description, we define $\mathbf{Z} \in \mathcal{Z} \subset \mathbb{R}^s$ as the complete confounder vector. The data generation process of these elements coincides with the graph in Figure 2(*a*). Empirically, the observational dataset can be denoted as $\{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{1 \leq i \leq n}$, where n is the sample size.

In many applications, the investigator can not collect all the confounders into the observed covariates $\mathbf{X}$. For the individual with complete confounder $\mathbf{Z}$, we denote the individual-level outcome of counterfactual treatment $\mathbf{T}$ by $Y_{\mathbf{Z}}(\mathbf{T})$ which is the ultimate goal of counterfactual prediction. We assume overlap, stable unit treatment value (SUTVA) (Rosenbaum and Rubin, 1983) are satisfied in this paper.

Because of confounding bias, directly applying supervised learning to estimate $\mathbb{E}[\mathbf{y}|\mathbf{X}, \mathbf{T}]$ is a biased estimation of the heterogeneous outcome conditional on covariates $\mathbb{E}[\mathbf{y}|\mathbf{X}, do(\mathbf{T})]$. Furthermore, even unbiased heterogeneous estimation $\mathbb{E}[\mathbf{y}|\mathbf{X}, do(\mathbf{T})]$ at a coarse granular level is still not accurate outcome prediction for individuals with confounders $\mathbf{Z}$ because of missing heterogeneity problem. Therefore, to step closer to individual-level prediction, it is necessary to recover the latent confounders to achieve finer granular heterogeneous outcome prediction by addressing confounding bias and missing heterogeneity.

4. Proposed Method

In this section, we introduce our proposed Factual Observation based Heterogeneity Learning (FOHL) algorithm in detail.

4.1. Distribution Modelling

From the graph shown in Figure 2(*a*), we can observe that given the latent confounders $\mathbf{Z}$, the treatments $\mathbf{T}$ and outcomes $\mathbf{y}$ are conditional independent of the covariates $\mathbf{X}$. Therefore, we model the joint distribution of these elements based on the probabilistic graph in Figure 2(*b*), which is Markov equivalent to Figure 2(*a*).

For implementation, we develop our model based on the architecture like IVAE (Khemakhem et al., 2020) to fit the empirical joint distribution of the dataset. The model architecture consists of three components, conditional prior component $p_\rho(\mathbf{Z}|\mathbf{X})$, encoder $q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{T},\mathbf{y})$ and decoder $p_\varphi(\mathbf{T},\mathbf{y}|\mathbf{Z})$. The components are designed as follows:

$$q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{T},\mathbf{y}) = \prod_{i=1}^{s'} q_\phi(\mathbf{z}_{,i}|\mathbf{X},\mathbf{T},\mathbf{y}) \quad , \quad p_\rho(\mathbf{Z}|\mathbf{X}) = \prod_{i=1}^{s'} p_\rho(\mathbf{z}_{,i}|\mathbf{X})$$
$$p_\varphi(\mathbf{T},\mathbf{y}|\mathbf{Z}) = p_\varphi(\mathbf{T}|\mathbf{Z})p_\varphi(\mathbf{y}|\mathbf{Z},\mathbf{T}) \quad , \quad p_\varphi(\mathbf{T}|\mathbf{Z}) = \prod_{i=1}^{d} p_\varphi(\mathbf{t}_{,i}|\mathbf{Z}),$$

where we set

$$q_\phi(\mathbf{z}_{,i}|\mathbf{X},\mathbf{T},\mathbf{y}) = \mathcal{N}(\mu_i^\phi(\mathbf{X},\mathbf{T},\mathbf{y}),(\sigma_i^\phi(\mathbf{X},\mathbf{T},\mathbf{y}))^2) \quad , \quad p_\rho(\mathbf{z}_{,i}|\mathbf{X}) = \mathcal{N}(\mu_i^\rho(\mathbf{X}),(\sigma_i^\rho(\mathbf{X}))^2),$$
$$p_\varphi(\mathbf{y}|\mathbf{Z},\mathbf{T}) = \mathcal{N}(\mu^\varphi(\mathbf{Z},\mathbf{T}),(\sigma^\varphi)^2) \quad , \quad p_\phi(\mathbf{t}_{,i}|\mathbf{Z}) = Bernoulli(\mathbf{g}_i^\phi(\mathbf{Z})).$$

The functions $\mu^{(\rho,\phi,\varphi)}(\cdot), \sigma^{(\rho,\phi)}(\cdot)$ and $\mathbf{g}^\phi(\cdot)$ are implemented by deep neural networks with parameters ρ,ϕ,φ. We set σ^φ as hyper-parameter. With the architecture above, we train the model to fit the distribution of the dataset by maximizing the evidence lower bound (ELBO) as follows:

$$\mathcal{L}_{elbo} = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{t}_i,y_i)} \left[\log p_\varphi(\mathbf{t}_i|\mathbf{z}) + \log p_\varphi(y_i|\mathbf{z},\mathbf{t}_i) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{t}_i,y_i)|p_\rho(\mathbf{z}|\mathbf{x}_i))\right]. \tag{1}$$

We can observe that the design of conditional prior component $p_\rho(\mathbf{Z}|\mathbf{X})$, makes our model compatible with the prediction of individuals outside the datasets. It enables us to infer the latent confounder representation based on only covariates and give the degenerate prediction at a coarse heterogeneity level which is the expected value of outcome conditional on the covariates.

We set covariates $\mathbf{X}$ as an ancestor of confounders $\mathbf{Z}$ rather than descendants for avoiding the intractable distribution assumption on covariates. In the wild application scenarios, the covariates can be continuous variables and mutually correlated of which the distribution is difficult to model. Furthermore, some noisy variables which are irrelevant to treatments and confounders may inevitably be collected into covariates (Zhang et al., 2021). However, the representation learned by our method, which maximizes the ELBO of $\prod_{i=1}^n p(\mathbf{t}_i,y_i|\mathbf{x}_i)$, does not reconstruct the covariates and will be less vulnerable to the noisy variables. This will be empirically demonstrated in the section on experiments.

4.2. Outcome Prediction

For the convenience of description, we define $Y(do(t^c)|\mathbf{x}, \mathbf{t}, y)$ as the predicted outcome of treatment $\mathbf{t}^c$ derived from our model for the individual with observed covariates $\mathbf{x}$, treatment $\mathbf{t}$ and outcome y. Our prediction process is composed of two steps.

At first, for the ith sample in the dataset, we can easily obtain the posterior distribution of the confounder representation with the encoder, $\hat{\mathbf{z}}_i \sim q_\phi(\mathbf{z}|\mathbf{x}_i, \mathbf{t}_i, y_i)$. Then given the posterior distribution of $\hat{\mathbf{z}}_i$, we can infer the outcome distribution of counterfactual treatment $\mathbf{t}^c$ with the decoder $p_\varphi(\mathbf{y}|\mathbf{Z}, \mathbf{T})$ as $\mathbb{E}_{\hat{\mathbf{z}}_i \sim q_\phi(\mathbf{z}|\mathbf{x}_i, \mathbf{t}_i, y_i)}\left[p_\varphi(\mathbf{y}|\hat{\mathbf{z}}_i, \mathbf{t}^c)\right]$. We estimate the expected value of the distribution as the prediction result, that is

$$Y(do(t^c)|\mathbf{x}_i, \mathbf{t}_i, y_i) = \mathbb{E}_{\hat{\mathbf{z}}_i \sim q_\phi(\mathbf{z}|\mathbf{x}_i, \mathbf{t}_i, y_i)}\left[\mu^\varphi(\mathbf{y}|\hat{\mathbf{z}}_i, \mathbf{t}^c)\right]. \tag{2}$$

Empirically, the estimation result can be approximated by repeatedly sampling $\hat{\mathbf{z}}_i^j \sim q_\phi(\mathbf{z}|\mathbf{x}_i, \mathbf{t}_i, y_i)$, $1 \leq j \leq m$ and calculate the following equation:

$$\hat{Y}(do(\mathbf{t}^c)|\mathbf{x}_i, \mathbf{t}_i, y_i) = \frac{1}{m}\sum_{j=1}^{m} \mu^\varphi(\hat{\mathbf{z}}_i^j, \mathbf{t}^c). \tag{3}$$

Besides, the trained model can also be applied to new individuals outside the datasets for degenerate prediction at a coarse heterogeneity level. Similarly, we define $Y(do(t^c)|\mathbf{x})$ as the prediction result conditional on only observed covariates $\mathbf{x}$. For adapting the method to this scenario, we instead sample the confounder representation from conditional prior distribution in the first step, formally $\hat{\mathbf{z}} \sim p_\rho(\mathbf{z}|\mathbf{x}_i)$. The second step in Equation 3 remains unchanged. The pseudo-code of the whole algorithm can be found in the appendix.

5. Theoretical Analysis

To show the validity of our strategy, we theoretically analyze the performance comparison of heterogeneous outcome prediction based on respectively the latent confounder representation learned from factual observations and only observed covariates under some mild conditions. More specifically, we assume that the distribution of latent confounder can be identified up to an invertible transformation as follows, which is also admitted by some previous works (Louizos et al., 2017; Miao et al., 2022).

Assumption 1 *There exists an invertible mapping f from the true latent confounder $\mathbf{z}$ to inferred latent confounder representation $\mathbf{z}'$. Formally, $p'(\mathbf{z}' = f(\mathbf{z}), \mathbf{x}, \mathbf{t}, y) \cdot |det(\frac{\partial f}{\partial \mathbf{z}})| = p(\mathbf{z}, \mathbf{x}, \mathbf{t}, y)$, where $det(\frac{\partial f}{\partial \mathbf{z}})$ is the determinant of Jacobian matrix of f, $p(\mathbf{z}, \mathbf{x}, \mathbf{t}, y)$ and $p'(\mathbf{z}', \mathbf{x}, \mathbf{t}, y)$ are respectively the true joint distribution and the distribution derived from our model.*

With this assumption, we can prove that our outcome prediction results are unbiased.

Proposition 1 *If the Assumption 1 holds, we have $Y(do(\mathbf{t}^c)|\mathbf{x}, \mathbf{t}, y) = \mathbb{E}[\mathbf{y}|\mathbf{x}, \mathbf{t}, y, do(\mathbf{t}^c)]$ and $Y(do(\mathbf{t}^c)|\mathbf{x}) = \mathbb{E}[\mathbf{y}|\mathbf{x}, do(\mathbf{t}^c)]$.*

Before presenting our theoretical results on missing heterogeneity, we make the following definition of counterfactual predictive error:

Definition 2 *We respectively define the predictive error based on the latent confounder representation and only covariates as follows:*

$$\begin{aligned}\mathcal{E}^{FOHL} &= \mathbb{E}_{\mathbf{z},\mathbf{x},\mathbf{t},y\sim p(\mathbf{z},\mathbf{x},\mathbf{t},\mathbf{y}),\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\left(Y_{\mathbf{z}}(\mathbf{t}^c)-Y(do(\mathbf{t}^c)|\mathbf{x},\mathbf{t},y)\right)^2\right],\\ \mathcal{E}^{\mathbf{X}} &= \mathbb{E}_{\mathbf{z},\mathbf{x}\sim p(\mathbf{z},\mathbf{x}),\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\left(Y_{\mathbf{z}}(\mathbf{t}^c)-Y(do(\mathbf{t}^c)|\mathbf{x})\right)^2\right],\end{aligned}$$

where $p^u(\mathbf{t})=\frac{1}{2^d}$ is the uniform distribution over the treatment space.

With the assumptions, definitions, and propositions above, we can prove the following proposition:

Proposition 3 *Assuming the individual outcome satisfies $Y_{\mathbf{z}}(\mathbf{t})=g(\mathbf{z},\mathbf{t})+\varepsilon$ where ε is a noise term with zero mean and σ^2 variance, $\mathcal{E}^{FOHL}$, $\mathcal{E}^{\mathbf{X}}$ can be written as following:*

$$\mathcal{E}^{FOHL} = \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x},\mathbf{t},y\sim p(\mathbf{x},\mathbf{t},y)}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right]+\sigma^2, \tag{4}$$

$$\mathcal{E}^{\mathbf{X}} = \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right]+\sigma^2. \tag{5}$$

According to the law of total variance, we have $\mathcal{E}^{FOHL}\leq\mathcal{E}^{\mathbf{X}}$.

Proposition 3 reveals the validity of heterogeneity learning from factual observations for addressing the missing heterogeneity problem. In this way, we deal with the outcome variation in the sub-population with the same covariates $\mathbf{X}$ and achieve finer granular heterogeneous outcome prediction (i.e. closer to individual-level). Detailed proof of the propositions above can be found in the appendix. Empirically, the superiority of our strategy can also be verified in practice. We show it in the section of experiments.

6. Experiments

In this section, we present our experimental results to show the effectiveness of our method. The evaluation requires the ground truth of counterfactual outcomes for individuals, which is not satisfied by the observational study in reality. Hence the experiments are conducted on synthetic datasets and semi-synthetic datasets.

6.1. Experimental Setup

We give a brief overview of the baselines and the evaluation metrics used in the experiments.

Baselines We compare our FOHL method with the methods listed below:

- Vanilla counterfactual prediction (Vcp): It directly applies feed-forward neural networks taking observed covariates and treatments as input to predict individual outcomes. The model is trained on the re-weighted dataset which removes the correlation between treatments and covariates. The sample weights are computed by density ratio estimation (Qin, 1998; Sugiyama et al., 2012; Bickel et al., 2007) as in Arbour et al. (2021).

- CEVAE (Louizos et al., 2017): It uses variational autoencoder (VAE) (Rezende et al., 2014; Kingma and Welling, 2013) to capture the joint distribution of latent confounders, observed covariates, treatments and outcomes. Although the original version is designed for conditional

average treatment effect estimation of binary treatment, we make some effort to extend this method for heterogeneous outcome prediction of multiple treatments by learning latent confounder representation from factual observations.

- DSCM (Pawlowski et al., 2020): It simply formulates the treatments and covariates as the ancestors of the outcome, and conducts counterfactual query on this SCM. We use the variational inference to model the structural equation as the amortised, explicit setting in Pawlowski et al. (2020).
- Deconfounder (Wang and Blei, 2019): It uses a factor model to compute the latent variables which can render the treatments conditionally independent as substitute confounders. Then it trained a predictive model taking substitute confounders and treatments as input. We choose VAE as the factor model which makes weaker assumptions about the data generation process.
- Deconfounder(+): It trained the predictive model which takes observed covariates, substitute confounders learned by Deconfounder and treatments as input.

Evaluation Metrics The methods are evaluated by the root mean square error (RMSE) of outcome estimation of the samples with all possible treatments in $\mathcal{T}$. Intuitively, the smaller RMSE implies that the estimation achieved by the method is closer to the individual-level counterfactual prediction, with a better recovery on the missing heterogeneity. Specifically, there are two evaluation settings in our experiments, which are within-sample setting and out-of-sample setting.

Under the within-sample setting, we evaluate methods on the samples in the observational dataset which is with factual observations. Latent confounder representation can be inferred by leveraging the previous treatment assignments and outcomes.

Under the out-of-sample setting, the evaluation is conducted on the samples out of the observational dataset which is in absence of previous observed treatments and outcomes. Since Deconfounder and Deconfounder(+) rely heavily on the assigned treatments to infer hidden confounders, they can not be easily adapted to the out-of-sample setting. Vcp keeps the same between the two settings. Therefore, only DSCM, CEVAE, and our FOHL method are applicable and meaningful to this setting. The performance comparison between the two settings can reveal the advantage of heterogeneity learning from factual observations to address missing heterogeneity problem.

6.2. Synthetic Datasets

We give a brief overview of how to generate the synthetic datasets and then present the experimental results on the datasets.

Datasets We generate the synthetic datasets in several steps. We generate the hidden confounders $\mathbf{Z}$, where the elements are sampled from a gaussian distribution $\mathbf{Z} \sim \mathcal{N}(0, \Sigma^z)$. Then for each sample in the dataset, we assign treatments based on the hidden confounders $\mathbf{Z}$. To be specific, we firstly compute the vector $\mathbf{L} = \mathbf{Z} \cdot \mathbf{A} + \varepsilon_L$, where $\mathbf{L}, \varepsilon_L \in \mathbb{R}^d, \mathbf{A} \in \mathbb{R}^{p \times d}$. The elements in $\mathbf{A}$ are independently generated from $\mathcal{N}(0, 0.8^2)$ and fixed for all the samples. The noise vector ϵ_L is generated once from $\mathcal{N}(0, 1.8^2)$ for each sample. Then the treatment variable $\mathbf{T}$ is calculated from $\mathbf{L}$. For each $j \in \{1, 2, ..., d\}$, if $\mathbf{l}_{,j} > 0$, we set $\mathbf{t}_{,j} = 1$, otherwise we set $\mathbf{t}_{,j} = 0$. The outcome is determined by both confounders $\mathbf{Z}$ and $\mathbf{T}$. We generate it from a pre-defined function: $\mathbf{y} = \sum_{j=1}^{p} \sum_{k=1}^{d} \mathbf{z}_{,j} d_{j,k} \mathbf{t}_{,k}$, where $\mathbf{D} \in \mathbb{R}^{p \times d}$ is a constant matrix. The observed covariates consist of two parts. One is the noisy measurement of (partial) hidden confounders $\mathbf{X}^m \in \mathbb{R}^{p_1}$. For each

Table 1: The experimental results on synthetic datasets. Mean and STD refer to the average value and standard deviation of the RMSE results in 10 times repeated experiments. Lower metrics means better performance. The best performance among the methods is marked bold.

Varying confounder measurement dimension p_1, Fixing noise scale $\sigma_x = 0.3$												
Within-Sample Setting												
p_1	$p_1 = 0$		$p_1 = 1$		$p_1 = 2$		$p_1 = 3$		$p_1 = 4$		$p_1 = 5$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	4.161	0.043	3.768	0.062	3.430	0.061	2.853	0.042	2.039	0.038	1.391	0.051
Deconfounder	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203
Deconfounder(+)	4.392	0.287	3.998	0.485	3.874	0.395	3.233	0.411	2.964	0.308	2.073	0.242
DSCM	3.784	0.022	3.265	0.029	3.023	0.018	2.536	0.034	1.787	0.026	1.190	0.024
CEVAE	2.107	0.234	2.016	0.041	1.844	0.031	2.283	0.058	1.588	0.021	1.551	0.092
FOHL	**1.920**	0.081	**1.862**	0.030	**1.722**	0.024	**1.703**	0.050	**1.449**	0.040	**1.172**	0.030
Out-of-Sample Setting												
DSCM	4.069	0.014	3.514	0.022	3.183	0.012	2.651	0.028	1.817	0.021	1.210	0.021
CEVAE	3.715	0.032	3.060	0.017	2.754	0.020	2.520	0.041	1.754	0.016	1.620	0.070
FOHL	3.732	0.022	3.071	0.013	2.740	0.015	2.300	0.028	1.662	0.032	1.189	0.027
Varying noise scale σ_x , Fixing confounder measurement dimension $p_1 = 4$												
Within-Sample Setting												
σ_x	$\sigma_x = 0.2$		$\sigma_x = 0.4$		$\sigma_x = 0.6$		$\sigma_x = 0.8$		$\sigma_x = 1.0$		$\sigma_x = 1.2$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	1.799	0.043	2.275	0.054	2.719	0.018	3.074	0.050	3.338	0.050	3.519	0.042
Deconfounder	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203	3.259	0.203
Deconfounder(+)	2.467	0.167	2.899	0.477	3.036	0.257	3.279	0.275	3.460	0.583	3.777	0.570
DSCM	1.611	0.027	1.988	0.025	2.386	0.022	2.710	0.027	2.950	0.027	3.143	0.031
CEVAE	1.534	0.030	1.682	0.031	1.800	0.031	1.869	0.027	1.895	0.034	1.883	0.035
FOHL	**1.384**	0.026	**1.555**	0.098	**1.678**	0.120	**1.724**	0.120	**1.751**	0.102	**1.755**	0.073
Out-of-Sample Setting												
DSCM	1.631	0.021	2.042	0.023	2.487	0.016	2.859	0.022	3.135	0.017	3.348	0.019
CEVAE	1.606	0.024	1.900	0.024	2.226	0.022	2.494	0.016	2.699	0.021	2.861	0.021
FOHL	1.524	0.024	1.848	0.059	2.210	0.047	2.506	0.042	2.734	0.030	2.921	0.027

$j \in \{1, 2, ..., p_1\}$, $\mathbf{x}_{,j} = \mathbf{z}_{,j} + \varepsilon_x$, where $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2)$ is measurement noise. The other part is noisy observations $\mathbf{X}^n \in \mathbb{R}^{p_2}$. The noisy observations are sampled from gaussian distribution $\mathbf{X}^n \sim \mathcal{N}(0, \Sigma^x)$ and irrelevant to treatments and outcomes. Finally, the observed covariates are the concatenate of the two parts. Formally, $\mathbf{X} = [\mathbf{X}^m, \mathbf{X}^n]$. Since only partial information of confounders is observed, it brings missing heterogeneity and confounding bias into the dataset.

For evaluation, we randomly assign new treatments to each sample by sampling treatments from uniform distribution: $\mathbf{t}_{,1}, \mathbf{t}_{,2}, ..., \mathbf{t}_{,d} \sim Bernoulli(0.5)$, and generate new outcomes. We calculate the RMSE of estimation of the new outcomes to compare different methods.

In these experiments, we set confounder dimension $p = 5$, treatment dimension $d = 20$, dimension of noisy observation $p_2 = 10$, and sample size $n = 10000$. More detailed information on the experimental setup is described in the appendix. We also conduct experiments under the setting where the latent confounders are generated from observed covariates. The detailed results are included in the appendix and show that our method still achieve superior performance.

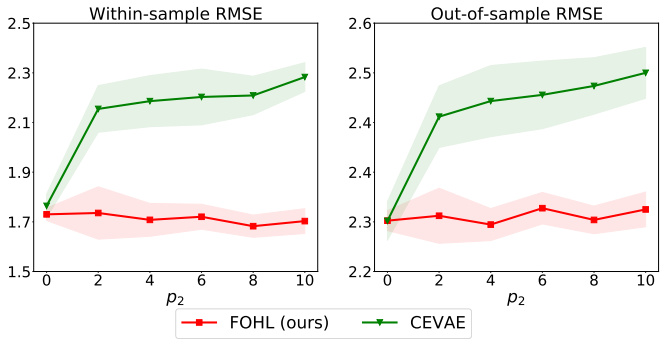


Figure 3: The influence of noisy measurement dimension p_2 on RMSE of prediction. We conduct experiments under the setting where $p_1 = 3$ and $\sigma_x = 0.3$. The shade region presents the interval [mean-std,mean+std] of RMSE.

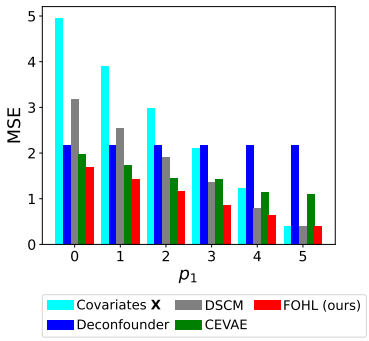


Figure 4: The confounder estimation MSE based on different methods. p_1 is confounder measurement dimension.

Results We conduct experiments under different settings by varying the dimension p_1 and noise scale of hidden confounder measurement σ_x. For each experiment setting, we repeatedly conduct the experiments for 10 times and calculate the mean value and standard deviation of RMSEs in the outcome estimation. The performance under both the within-sample setting and the out-of-sample setting is reported in Table 1.

From the results in Table 1, we can observe that when less confounder information is contained in the observed covariates (e.g. small p_1 and large σ_x), Vcp performs worst among the different methods because it suffers from the confounding bias and missing heterogeneity problem due to the lack of confounder information in the data. Deconfounder can infer the hidden confounders when treatments are available and reduce RMSE based on Vcp to some degree when few individual information are measured. Deconfounder(+) roughly concatenates the inferred representation and noisy observations, therefore achieving unsatisfactory performance. DSCM assumes the latent confounders are marginally independent of observed covariates and treatments. Hence it suffers from the confounding bias and performs poorly under different settings. CEVAE and our FOHL method both learn the latent confounder representation from the factual observations to achieve finer granular heterogeneous prediction. The counterfactual prediction results are more accurate for individuals under the within-sample setting. The performance comparison between within-sample setting and out-of-sample setting highlights the advantage of addressing missing heterogeneity for finer granular heterogeneous outcome prediction. Since CEVAE learns the latent confounder representation to reconstruct observed covariates, it is vulnerable to noisy observations in the covariates. This has also been studied by previous literature (Rissanen and Marttinen, 2021). Therefore, our method achieves the best performance among all the methods.

We conduct confounder estimation experiments with $\sigma_x = 0.3$, which trains a neural network to predict the true underlying confounders based on the learned representations of different methods or only covariates. The results shown in Figure 4 reveal that our method which simultaneously leverages covariates, treatments, and outcomes, achieves the smallest estimation error and performs best in recovering the underlying confounders. We also conduct experiments to predict the noisy

Table 2: The experimental results of click probability prediction on semi-synthetic datasets ($\times 10^{-1}$).

Varying confounder measurement dimension p_1, Fixing noise scale $\sigma_x = 0.5$												
Within-Sample Setting												
p_1	$p_1 = 0$		$p_1 = 1$		$p_1 = 2$		$p_1 = 3$		$p_1 = 4$		$p_1 = 5$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	1.779	0.014	1.629	0.036	1.539	0.020	1.502	0.026	1.258	0.022	1.181	0.052
Deconfounder	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117
Deconfounder(+)	1.443	0.325	1.380	0.273	1.379	0.243	1.165	0.103	1.126	0.285	1.117	0.222
DSCM	1.633	0.007	1.511	0.017	1.414	0.013	1.408	0.041	1.267	0.066	1.168	0.063
CEVAE	1.326	0.200	1.185	0.044	1.125	0.046	1.162	0.037	1.083	0.030	1.049	0.039
FOHL	**1.115**	0.224	**1.037**	0.101	**1.051**	0.086	**1.003**	0.104	**0.934**	0.059	**0.830**	0.035
Out-of-Sample Setting												
DSCM	1.788	0.008	1.651	0.011	1.545	0.018	1.517	0.028	1.308	0.043	1.214	0.042
CEVAE	1.778	0.046	1.467	0.039	1.362	0.018	1.271	0.040	1.105	0.040	1.063	0.042
FOHL	1.696	0.016	1.437	0.014	1.333	0.017	1.245	0.026	1.047	0.044	0.952	0.023
Varying noise scale σ_x , Fixing confounder measurement dimension $p_1 = 4$												
Within-Sample Setting												
σ_x	$\sigma_x = 0.2$		$\sigma_x = 0.4$		$\sigma_x = 0.6$		$\sigma_x = 0.8$		$\sigma_x = 1.0$		$\sigma_x = 1.2$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	0.700	0.032	1.105	0.033	1.385	0.019	1.524	0.019	1.616	0.020	1.668	0.017
Deconfounder	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117	1.450	0.117
Deconfounder(+)	0.916	0.247	1.054	0.153	1.251	0.274	1.353	0.248	1.296	0.164	1.348	0.528
DSCM	0.690	0.016	1.120	0.067	1.325	0.082	1.453	0.072	1.545	0.064	1.555	0.023
CEVAE	1.029	0.098	1.046	0.035	1.084	0.032	1.107	0.040	1.127	0.038	1.147	0.032
FOHL	**0.681**	0.017	**0.940**	0.046	**0.906**	0.052	**0.902**	0.046	**0.929**	0.077	**0.968**	0.101
Out-of-Sample Setting												
DSCM	0.705	0.016	1.156	0.052	1.393	0.054	1.542	0.049	1.644	0.046	1.670	0.030
CEVAE	1.077	0.074	1.060	0.036	1.147	0.036	1.238	0.056	1.335	0.054	1.408	0.058
FOHL	0.692	0.014	1.004	0.038	1.075	0.036	1.168	0.025	1.260	0.037	1.343	0.040

observations $\mathbf{X}^n$ based on the representation learned by FOHL and CEVAE. The prediction MSE of representation learned by CEVAE and FOHL are respectively around 1.800 and 9.800 under various settings. This shows that our method can resist the influence of noise in the covariates while CEVAE suffers from it. We empirically investigate the influence of the noisy measurement dimension on performance. The results in Figure 3 show that the performance of our method is overall stable w.r.t the quantity of noisy measurement in the observed covariates. However, CEVAE is vulnerable to it.

We explore the influence of latent variable dimension and hyper-parameter σ^{φ} in the model on the performance. The results of these experiments can be found in the appendix.

6.3. Semi-synthetic Datasets

We conduct some experiments on the semi-synthetic datasets generated by Recsim (Ie et al., 2019) which approximates the recommendation scenarios.

Datasets There is an environment[1] simulating document recommendation in Recsim. Each document is depicted by its category and quality. The confounder of a user is a vector of affinity

1. https://github.com/google-research/recsim/blob/master/recsim/environments/interest_exploration.py

to each document category $\mathbf{Z} \in \mathbb{R}^p$, where p is the number of categories. The recommended items form the treatment vector $\mathbf{T} \in \{0, 1\}^d$, where d is the number of documents in the pool and each bit means whether the corresponding document is recommended.

To generate the observational dataset, we assign treatments in a similar manner to that of synthetic datasets. Given the user confounders and recommended documents (i.e. treatments), the simulator can generate the click probability on the recommended document bundle, which is treated as the outcome. We also concatenate the noisy measurement of user confounders and noisy observations as the observed covariates. For evaluation, we calculate the RMSE of prediction in the testing dataset, where each document is recommended independently with 50% probability. Due to the space limitations of the main paper, we leave a detailed description of the experimental setup in the appendix.

Results We conduct experiments under different settings by varying the confounder measurement dimension and noise scale of confounder measurement. The results are shown in Table 2.

From the results of the semi-synthetic dataset, we can observe a similar trend to that of experiments on synthetic datasets. When the confounder information in the observed covariates is little, the performance of FOHL and CEVAE under the within-sample setting is superior to the counterpart under the out-of-sample setting. This coincides with the intuition and indicates the benefit of leveraging factual observations to heterogeneity learning for finer granular heterogeneous outcome prediction, especially when little individual confounder information is observed in covariates.

7. Conclusion

In this paper, we investigate the problem of finer granular heterogeneous outcome prediction under the setting of missing confounders which is closer to individual-level counterfactual prediction. We propose a Factual Observation based Heterogeneity Learning (FOHL) method to simultaneously address the confounding bias and missing heterogeneity problem. Theoretical analysis reveals the advantage of our strategy. Extensive experimental results show the effectiveness of our method.

Acknowledgments

Peng Cui's research was supported in part by National Key R&D Program of China (No.2018AAA0102004, No. 2020AAA0106300), National Natural Science Foundation of China (No.U1936219, 62141607), Beijing Academy of Artificial Intelligence (BAAI). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131, 72133002); the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.

References

David Arbour, Drew Dimmery, and Arjun Sondhi. Permutation weighting. In *International Conference on Machine Learning*, pages 331–341. PMLR, 2021.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020a.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020b.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

Issa J Dahabreh, Rodney Hayward, and David M Kent. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, 45(6):2184–2193, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

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

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

James Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, pages 441–462, 1997.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

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

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*, pages 2207–2217. PMLR, 2020.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, and Mihaela van der Schaar. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:2282–2292, 2020.

Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934, 2015.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Wang Miao and Eric Tchetgen Tchetgen. Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *American journal of epidemiology*, 185(10): 950–953, 2017.

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.

Wang Miao, Wenjie Hu, Elizabeth L. Ogburn, and Xiaohua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 2022.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33: 857–869, 2020.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a.

Judea Pearl. *Causality*. Cambridge university press, 2009b.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34, 2021.

Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. In *First Conference on Causal Learning and Reasoning*, 2021.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

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

Can Xu, Ahmed Alaa, Ioana Bica, Brent Ershoff, Maxime Cannesson, and Mihaela van der Schaar. Learning matching representations for individualized organ transplantation allocation. In *International Conference on Artificial Intelligence and Statistics*, pages 2134–2142. PMLR, 2021.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 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.

Shuxi Zeng, Serge Assaad, Chenyang Tao, Shounak Datta, Lawrence Carin, and Fan Li. Double robust representation learning for counterfactual prediction. *arXiv preprint arXiv:2010.07866*, 2020.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *AAAI*, 2021.

Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33:19705–19715, 2020.

Appendix A. Proofs

Proposition 4 ***(Restated)*** *If the Assumption 1 holds, we have* $Y(do(\mathbf{t}^c)|\mathbf{x},\mathbf{t},y) = \mathbb{E}[\mathbf{y}|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c)]$ *and* $Y(do(\mathbf{t}^c)|\mathbf{x}) = \mathbb{E}[\mathbf{y}|\mathbf{x},do(\mathbf{t}^c)]$.

Proof According to Assumption 1, we have:

$$\begin{aligned}
p'(\mathbf{x},\mathbf{t},y) &= \int_{\mathbf{z}'} p'(\mathbf{z}',\mathbf{x},\mathbf{t},y)d\mathbf{z}' = \int_{\mathbf{z}'} \frac{p(f^{-1}(\mathbf{z}'),\mathbf{x},\mathbf{t},y)}{|det(\frac{\partial f}{\partial \mathbf{z}})|_{\mathbf{z}=f^{-1}(\mathbf{z}')}} d\mathbf{z}' \\
&= \int_{\mathbf{z}} \frac{p(\mathbf{z},\mathbf{x},\mathbf{t},y)}{|det(\frac{\partial f}{\partial \mathbf{z}})|} \cdot |det(\frac{\partial f}{\partial \mathbf{z}})|d\mathbf{z} = \int_{\mathbf{z}} p(\mathbf{z},\mathbf{x},\mathbf{t},y)d\mathbf{z} = p(\mathbf{x},\mathbf{t},y). \\
p(\mathbf{t},y,\mathbf{z}) &= \int_{\mathbf{x}} p(\mathbf{x},\mathbf{t},y,\mathbf{z})d\mathbf{x} = \int_{\mathbf{x}} p'(\mathbf{x},\mathbf{t},y,f(\mathbf{z}))|det(\frac{\partial f}{\partial \mathbf{z}})|d\mathbf{x} && (6) \\
&= p'(\mathbf{t},y,f(\mathbf{z}))|det(\frac{\partial f}{\partial \mathbf{z}})|. && (7) \\
p(\mathbf{t},\mathbf{z}) &= p'(\mathbf{t},f(\mathbf{z}))|det(\frac{\partial f}{\partial \mathbf{z}})|. && (8) \\
p(\mathbf{x},\mathbf{z}) &= p'(\mathbf{x},f(\mathbf{z}))|det(\frac{\partial f}{\partial \mathbf{z}})|. && (9)
\end{aligned}$$

Therefore, denoting $\mathbf{z}'$ as the latent representation derived from our model and $p'(\mathbf{y}'|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c))$ as the outcome distribution with intervening $\mathbf{t}^c$ given observed $\mathbf{x}$, $\mathbf{t}$ and y, we can conclude that

$$\begin{aligned}
p'(\mathbf{y}'|\mathbf{z}',\mathbf{t}^c) &= \frac{p'(\mathbf{y}',\mathbf{z}',\mathbf{t}^c)}{p'(\mathbf{z}',\mathbf{t}^c)} = p(\mathbf{y}'|f^{-1}(\mathbf{z}'),\mathbf{t}^c) && (10) \\
p'(\mathbf{z}'|\mathbf{x},\mathbf{t},y) &= \frac{p'(\mathbf{z}',\mathbf{x},\mathbf{t},y)}{p'(\mathbf{x},\mathbf{t},y)} = \frac{p(f^{-1}(\mathbf{z}')|\mathbf{x},\mathbf{t},y)}{|det(\frac{\partial f}{\partial \mathbf{z}})|_{\mathbf{z}=f^{-1}(\mathbf{z}')}} && (11) \\
p'(\mathbf{y}'|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c)) &= \int_{\mathbf{z}'} p'(\mathbf{z}'|\mathbf{x},\mathbf{t},y)p'(\mathbf{y}'|\mathbf{z}',\mathbf{t}^c)d\mathbf{z}' && (12) \\
&= \int_{\mathbf{z}'} \frac{p(f^{-1}(\mathbf{z}')|\mathbf{x},\mathbf{t},y)}{|det(\frac{\partial f}{\partial \mathbf{z}})|_{\mathbf{z}=f^{-1}(\mathbf{z}')}} p(\mathbf{y}'|f^{-1}(\mathbf{z}'),\mathbf{t}^c)d\mathbf{z}' && (13) \\
&= \int_{\mathbf{z}} \frac{p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}{|det(\frac{\partial f}{\partial \mathbf{z}})|} p(\mathbf{y}'|\mathbf{z},\mathbf{t}^c)df(\mathbf{z}) && (14) \\
&= \int_{\mathbf{z}} p(\mathbf{z}|\mathbf{x},\mathbf{t},y)p(\mathbf{y}'|\mathbf{z},\mathbf{t}^c)d\mathbf{z} = p(\mathbf{y}'|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c)) && (15)
\end{aligned}$$

Then we can have $Y(do(\mathbf{t}^c)|\mathbf{x},\mathbf{t},y) = \mathbb{E}_{\mathbf{y}\sim p'(\mathbf{y}|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c))}[\mathbf{y}] = \mathbb{E}[\mathbf{y}|\mathbf{x},\mathbf{t},y,do(\mathbf{t}^c)]$. ■

Proposition 5 ***(Restated)*** *Assuming the individual outcome satisfies* $Y_{\mathbf{z}}(\mathbf{t}) = g(\mathbf{z},\mathbf{t}) + \varepsilon$ *where* ε *is a noise term with zero mean and* σ^2 *variance,* $\mathcal{E}^{FOHL}$, $\mathcal{E}^{\mathbf{X}}$ *can be written as following:*

$$\begin{aligned}
\mathcal{E}^{FOHL} &= \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x},\mathbf{t},y\sim p(\mathbf{x},\mathbf{t},y)}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}\left[g(\mathbf{z},\mathbf{t}^c|)\right]\right]\right] + \sigma^2, && (16) \\
\mathcal{E}^{\mathbf{X}} &= \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right] + \sigma^2. && (17)
\end{aligned}$$

According to the law of total variance, we have $\mathcal{E}^{FOHL} \leq \mathcal{E}^{\mathbf{X}}$.

Proof

$$
\begin{aligned}
\mathcal{E}^{FOHL} &= \int_{\mathbf{t}^c} p^u(\mathbf{t}^c) \int_{\mathbf{x},\mathbf{t},y,\mathbf{z}} p(\mathbf{x},\mathbf{t},y)p(\mathbf{z}|\mathbf{x},\mathbf{t},y)[\mathbb{E}[(Y_{\mathbf{z}}(\mathbf{t}^c) - \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}[g(\mathbf{z},\mathbf{t}^c)])^2]]d\mathbf{x}d\mathbf{t}dyd\mathbf{z}d\mathbf{t}^c \\
&= \int_{\mathbf{t}^c} p^u(\mathbf{t}^c) \int_{\mathbf{x},\mathbf{t},y} p(\mathbf{x},\mathbf{t},y)[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}[g(\mathbf{z},\mathbf{t}^c)] + \sigma^2]d\mathbf{x}d\mathbf{t}d\mathbf{z}d\mathbf{t}^c \\
&= \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x},\mathbf{t},y\sim p(\mathbf{x},\mathbf{t},y)}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right] + \sigma^2 \\
&= \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbb{E}_{\mathbf{t},y\sim p(\mathbf{t},y|\mathbf{x})}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right]\right] + \sigma^2 \qquad (18)
\end{aligned}
$$

Similarly, we can also have

$$\mathcal{E}^{\mathbf{X}} = \mathbb{E}_{\mathbf{t}^c\sim p^u(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right]\right] + \sigma^2.$$

According to the law of total variance, we have

$$Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left[g(\mathbf{z},\mathbf{t}^c)\right] \geq \mathbb{E}_{\mathbf{t},y\sim p(\mathbf{t},y|\mathbf{x})}\left[Var_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}\left[g(\mathbf{z},\mathbf{t}^c)\right]\right], \forall \mathbf{t}^c \in \mathcal{T}, \mathbf{x} \in \mathcal{X}.$$

The equality only holds when $\forall \mathbf{x}, \mathbf{t}, y, \mathbf{t}^c, \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}[g(\mathbf{z},\mathbf{t}^c)] = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x},\mathbf{t},y)}[g(\mathbf{z},\mathbf{t}^c)]$. Therefore, we have $\mathcal{E}^{FOHL} \leq \mathcal{E}^{\mathbf{X}}$.

■

Appendix B. Pseudo-code of FOHL

The pseudo-code can be found in Algorithm 1.

Appendix C. Experimental details

In the synthetic datasets, the matrix $\Sigma^z \in \mathbb{R}^{p\times p}$. For $i \neq j$, $\Sigma^z_{i,j} = 0.2$ and for $i = j$, $\Sigma^z_{i,j} = 1.0$. The matrix $\Sigma^x \in \mathbb{R}^{p\times p}$. For $i \neq j$, $\Sigma^x_{i,j} = 0.8$ and for $i = j$, $\Sigma^x_{i,j} = 1.0$. The matrix $\mathbf{D}$ is generated from gaussian distribution, $d_{i,j} = a_{i,j}/4 + \mathcal{N}(0, 0.1^2) + 0.1$.

In the semi-synthetic datasets, the user confounder $\mathbf{Z}$ is also sampled from gaussian distribution $\mathcal{N}(0, \Sigma^z)$. For $i \neq j$, $\Sigma^z_{i,j} = 0.3$ and for $i = j$, $\Sigma^z_{i,j} = 1.0$.

The observed covariates in the semi-synthetic datasets also consists of two parts. One is the noisy measurement of hidden confounders $\mathbf{X}^m \in \mathbb{R}^{p_1}$. For each $j \in \{1, 2, ..., p_1\}$, $\mathbf{x}_{,j} = \mathbf{z}_{,j} + \varepsilon_x$, where $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2)$ is measurement noise. The other part is noisy observations $\mathbf{X}^n \in \mathbb{R}^{p_2}$. The noisy observations are sampled from guassian distribution $\mathbf{X}^n \sim \mathcal{N}(0, \Sigma^x)$. Finally, the observed covariates is the concatenate of the two parts. Formally, $\mathbf{X} = [\mathbf{X}^m, \mathbf{X}^n]$. $\Sigma^x = Diag(\Sigma^{(1)}, \Sigma^{(2)}, \Sigma^{(3)})$. For $k \in \{1, 2, 3\}$, $\Sigma^{(k)} \in \mathbb{R}^{\frac{p_2}{3}\times\frac{p_2}{3}}$. For $i \neq j$, $\Sigma^{(k)}_{i,j} = 0.85$ and for $i = j$, $\Sigma^{(k)}_{i,j} = 1.0$. We set the sample size $n = 10000$, the dimension of confounders $p = 5$, the dimension treatment $d = 50$, the dimension of noisy observations $p_2 = 15$.

The encoder, decoder and conditional prior components are implemented by a neural networks with two hidden layers of size 50. The learning rate is set to be 10^{-3}. We use the ELU function as activation functions. The model is trained by 3000 epochs using Adam optimizer. The hyper-parameter $(\sigma^{\varphi})^2 = 1.0$ in the experiments of synthetic datasets and $(\sigma^{\varphi})^2 = \frac{1}{800}$ in the experiments of semi-synthetic datasets.

Algorithm 1 Factual Observation based Heterogeneity Learning (FOHL)

Input: Observational data $\{(\mathbf{x}_j, \mathbf{t}_j, y_j)\}_{1 \le j \le n}$, learning rate λ_p, new treatments for the ith sample under within-sample setting $\mathbf{t}^{in}$, new samples and new treatments $\left(\mathbf{x}^{out}, \mathbf{t}^{out}\right)$ under out-of-sample setting.

Output: Predicted treatment outcome y^{in} and y^{out}.

Train the model, including encoder $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{T}, \mathbf{y})$, decoder $p_\varphi(\mathbf{T}|\mathbf{Z})$, $p_\varphi(\mathbf{y}|\mathbf{Z}, \mathbf{T})$ and conditional prior $p_\rho(\mathbf{Z}|\mathbf{X})$.

Set $y^{in} \leftarrow 0$. // Under within-sample setting.

for $k = 1, 2, ..., m$ **do**

  Sample $r \sim \mathcal{N}(0, I)$.

  Compute $\hat{\mathbf{z}}^{in} \leftarrow \mu^\phi(\mathbf{x}_i, \mathbf{t}_i, y_i) + r \odot \sigma^\phi(\mathbf{x}_i, \mathbf{t}_i, y_i)$.

  Update $y^{in} \leftarrow y^{in} + \frac{1}{m} \cdot \mu^\varphi(\hat{\mathbf{z}}^{in}, \mathbf{t}^{in})$.

end

Set $y^{out} \leftarrow 0$. // Under out-of-sample setting.

for $k = 1, 2, ..., m$ **do**

  Sample $r \sim \mathcal{N}(0, I)$.

  Compute $\hat{\mathbf{z}}^{out} \leftarrow \mu^\rho(\mathbf{x}^{out}) + r \odot \sigma^\rho(\mathbf{x}^{out})$

  Update $y^{out} \leftarrow y^{out} + \frac{1}{m} \cdot \mu^\varphi(\hat{\mathbf{z}}^{out}, \mathbf{t}^{out})$.

end

Return predicted outcome y^{in} and y^{out}.

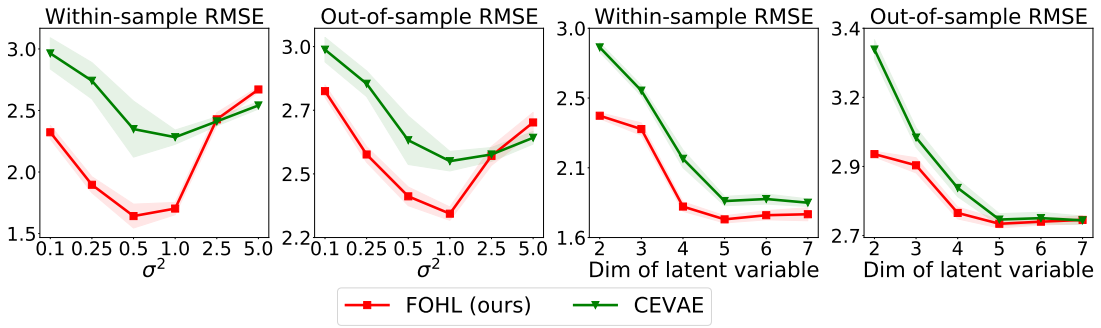


Figure 5: The influence of σ^φ and latent variable dimension on RMSE of counterfactual prediction. We conduct experiments under the original synthetic dataset of main paper where $p_1 = 3$ and $\sigma_x = 0.3$ for analysis on σ^φ and $p_1 = 2$ and $\sigma_x = 0.3$ for analysis on latent variable dimension.

Appendix D. Parameter Analysis

We conduct parameter analysis on synthetic datasets to show the characteristic of our method. To show the influence of $(\sigma^\varphi)^2$ and the latent variable dimension in model on the performance, we plot the curve of prediction error results in Figure 5. The results reflect that the performance of FOHL and CEVAE is stable w.r.t to the latent variable dimension. When the dimension is excessively small, the performance significantly declines.

Table 3: The experimental results on synthetic datasets where confounders are generated from covariates.

Varying observed dimension p_1, Fixing $\sigma_z = 0.3$												
Within-Sample Setting												
p_1	$p_1 = 0$		$p_1 = 1$		$p_1 = 2$		$p_1 = 3$		$p_1 = 4$		$p_1 = 5$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	4.016	0.041	3.938	0.042	3.652	0.044	3.065	0.061	2.789	0.070	1.496	0.045
Deconfounder	3.393	0.287	3.393	0.287	3.393	0.287	3.393	0.287	3.393	0.287	3.393	0.287
Deconfounder(+)	3.614	0.284	3.862	0.306	3.793	0.310	3.287	0.225	3.657	0.357	2.529	0.208
DSCM	3.674	0.028	3.452	0.019	3.104	0.032	2.653	0.046	2.430	0.020	1.330	0.054
CEVAE	2.468	0.254	2.358	0.094	2.542	0.256	2.372	0.033	2.946	0.269	1.495	0.037
FOHL	**2.198**	0.066	**2.191**	0.096	**2.173**	0.082	**1.946**	0.071	**1.746**	0.043	**1.337**	0.046
Out-of-Sample Setting												
DSCM	3.880	0.011	3.514	0.022	3.338	0.021	2.743	0.037	2.487	0.023	1.342	0.051
CEVAE	3.612	0.038	3.450	0.055	3.286	0.307	2.683	0.015	2.988	0.144	1.515	0.033
FOHL	3.695	0.036	3.494	0.033	3.026	0.030	2.600	0.032	2.321	0.033	1.347	0.042
Varying σ_z , Fixing $p_1 = 3$												
Within-Sample Setting												
σ_x	$\sigma_x = 0.2$		$\sigma_x = 0.4$		$\sigma_x = 0.6$		$\sigma_x = 0.8$		$\sigma_x = 1.0$		$\sigma_x = 1.2$	
Methods	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Vcp	2.902	0.044	3.233	0.057	3.633	0.038	4.068	0.048	4.603	0.046	5.071	0.039
Deconfounder	3.694	0.322	3.797	0.280	3.742	0.249	3.809	0.387	4.389	0.225	4.751	0.595
Deconfounder(+)	3.157	0.221	3.603	0.471	3.753	0.412	3.933	0.533	4.286	0.607	4.721	0.494
DSCM	2.573	0.045	2.786	0.044	3.129	0.043	3.542	0.041	4.027	0.029	4.536	0.037
CEVAE	2.435	0.080	2.422	0.080	2.709	0.211	3.148	0.385	3.459	0.320	3.857	0.278
FOHL	**1.897**	0.063	**2.040**	0.062	**2.303**	0.058	**2.655**	0.053	**3.206**	0.086	**3.784**	0.109
Out-of-Sample Setting												
DSCM	2.619	0.037	2.912	0.034	3.330	0.024	3.778	0.022	4.286	0.014	4.807	0.024
CEVAE	2.599	0.037	2.816	0.040	3.203	0.109	3.657	0.134	4.069	0.110	4.509	0.103
FOHL	2.453	0.031	2.782	0.033	3.191	0.037	3.669	0.023	4.274	0.052	4.921	0.067

Appendix E. Results of Experiments where covariates point towards confounders

In many scenarios, the latent confounders may be generated from covariates. For example, the salary level is the privacy of many people while their jobs can be observed. To verify the effectiveness of our method under this setting, we set $\mathbf{X}^t \sim \mathcal{N}(0, \Sigma^z)$ where Σ^z keeps same as in the main paper. The latent confounders are generated by $\mathbf{z}_{,j} = \mathbf{x}_{,j} + \varepsilon_z, \varepsilon_z \sim \mathcal{N}(0, \sigma_z^2)$. The observed covariates consists of two parts, $\mathbf{X} = [\mathbf{X}^m, \mathbf{X}^n]$. $\mathbf{X}^m \in \mathbb{R}^{p_1}$ consists of p_1 dimension of $\mathbf{X}^m$ while $\mathbf{X}^n$ keeps same as in the main paper. The generation of the other elements is the same to the setting in the main paper. The results are reported in.