Targeted VAE: Variational and Targeted Learning for Causal Inference

Matthew J. Vowels
CVSSP
University of Surrey
Guildford, U.K.
m.j.vowels@surrey.ac.uk

Necati Cihan Camgoz
CVSSP
University of Surrey
Guildford, U.K.
n.camgoz@surrey.ac.uk

Richard Bowden
CVSSP
University of Surrey
Guildford, U.K.
r.bowden@surrey.ac.uk

***Abstract*—Undertaking causal inference with observational data is incredibly useful across a wide range of tasks including the development of medical treatments, advertisements and marketing, and policy making. There are two significant challenges associated with undertaking causal inference using observational data: treatment assignment heterogeneity (*i.e.*, differences between the treated and untreated groups), and an absence of counterfactual data (*i.e.*, not knowing what would have happened if an individual who did get treatment, were instead to have not been treated). We address these two challenges by combining structured inference and targeted learning. In terms of structure, we factorize the joint distribution into risk, confounding, instrumental, and miscellaneous factors, and in terms of targeted learning, we apply a regularizer derived from the influence curve in order to reduce residual bias. An ablation study is undertaken, and an evaluation on benchmark datasets demonstrates that TVAE has competitive and state of the art performance across.**

***Index Terms*—causal inference, targeted learning, variational inference, machine learning**

I. Introduction

The estimation of the causal effects of interventions or treatments on outcomes is of the upmost importance across a range of decision making processes, such as policy making [1], advertisement [2], the development of medical treatments [3], and the evaluation of evidence within legal frameworks [4], [5]. Despite the common preference for Randomized Controlled Trial (RCT) data over observational data, this preference is not always justified. Besides the lower cost and fewer ethical concerns, observational data may provide a number of statistical advantages including greater statistical power and increased generalizability [6]. However, there are two main challenges when dealing with observational data. Firstly, the group that receives treatment is usually not equivalent to the group that does not (treatment assignment heterogeneity), resulting in selection bias and confounding due to associated covariates. For example, young people may prefer surgery, older people may prefer medication. Secondly, we are unable to directly estimate the causal effect of treatment, because only the factual outcome for a given treatment assignment is available. In other words, we do not have the counterfactual associated with the outcome for a different treatment assignment to that which was given. Treatment effect inference with observational data is concerned with finding ways to estimate the causal effect by considering the expected differences between factual and counterfactual outcomes.

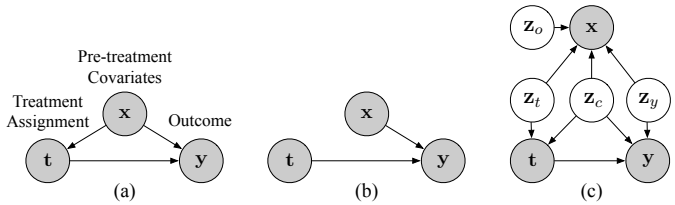

Fig. 1. Directed Acyclic Graphs (DAGs) for (a) the problem of estimating the effect of treatment $\mathbf{t}$ on outcome $\mathbf{y}$ with confounding $\mathbf{x}$. DAG (b) reflects an RCT. DAG (c) illustrates TVAE and is an extension of the DAG in [11], where the structure is *a priori* assumed to factorize into into risk $\mathbf{z}_y$, instrumental $\mathbf{z}_t$, and confounding factors $\mathbf{z}_c$. We extend their model with $\mathbf{z}_o$ to account for that fact that not all covariates will be related to treatment and/or outcome.

We seek to address the two challenges by proposing a method that enables the estimation of causal effects from observational data by leveraging techniques from the targeted learning literature. Specifically, Targeted Maximum Likelihood Estimation (TMLE) yields asymptotically efficient, unbiased, and doubly robust estimation of the causal effect, making it attractive as a causal inference method in its own right [7]–[10]. We incorporate targeted learning into a variational latent model, trained according to the approximate maximum likelihood paradigm. Doing so enables us to infer hidden confounders from proxy variables in the dataset, and to estimate average treatment effects, as well as conditional treatment effects. Estimating the latter is especially important for treatments that interact with patient attributes, whilst also being crucial for facilitating individualized treatment assignment. Thus, we propose the Targeted Variational AutoEncoder (TVAE), undertake an ablation study and also compare our method's performance against alternatives on two benchmark datasets.[1]

Accepted to Causal Learning (Special Track) at 2021 IEEE International Conference on Smart Data Services (SMDS).

[1]Code is available at https://github.com/matthewvowels1/TVAE_release

II. Background

Problem Formulation: A characterization of the problem of causal inference with no unobserved confounders is depicted in the Directed Acyclic Graphs (DAGs) shown in Figs. 1(a) and 1(b). For an accessible overview of the relevant background concerning causal inference and graphs, consider [12]–[14]. Fig. 1(a) is characteristic of observational data, where the assignment of treatment is related to the covariates. Fig. 1(b) is characteristic of the ideal RCT, where the treatment is unrelated to the covariates. Here, $\mathbf{x}_i \sim p(\mathbf{x}) \in \mathbb{R}^m$ represents the m-dimensional, pre-treatment covariates for individual i assigned factual treatment $t_i \sim p(\mathbf{t}|\mathbf{x})$ resulting in outcome $y_i \sim p(\mathbf{y}|\mathbf{x}, \mathbf{t})$. Together, these constitute dataset $\mathcal{D} = \{[y_i, t_i, \mathbf{x}_i]\}_{i=1}^N$ where N is the sample size. We occasionally refer to a *potential outcome* [15] for a particular treatment as $y(t)$, where the value of t may or may not correspond to that which has been observed for a particular individual (*i.e.*, it is a counterfactual outcome).

The conditional average treatment effect for an individual with covariates $\mathbf{x}_i$ may be defined as $\tau_i(\mathbf{x}_i) = \mathbb{E}[y_i|\mathbf{x}_i, \text{do}(\mathbf{t} = 1) - y_i|\mathbf{x}_i, \text{do}(\mathbf{t} = 0)]$, where the expectation accounts for the non-determinism of the outcome [16]. Alternatively, by comparing the post-intervention distributions when we intervene on treatment $\mathbf{t}$, the Average Treatment Effect (ATE) is $\tau(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\mathbb{E}[\mathbf{y}|\mathbf{x}, \text{do}(\mathbf{t} = 1)] - \mathbb{E}[\mathbf{y}|\mathbf{x}, \text{do}(\mathbf{t} = 0)]]$. Here, $\text{do}(\mathbf{t})$ indicates the intervention on $\mathbf{t}$, setting it to a prescribed static value, dynamic value, or distribution and therefore removing any dependencies it originally had [4], [7], [8]. This scenario corresponds with Fig. 1(b), where treatment $\mathbf{t}$ is no longer a function of the covariates $\mathbf{x}$.

Using an estimator[2] for the conditional mean $Q(\mathbf{t}, \mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{t}, \mathbf{x})$, we can calculate the Average Treatment Effect (ATE) and the empirical error for estimation of the ATE (eATE).[3] The estimated Average Treatment Effect (ATE) and error on the estimation of ATE (eATE) are given in Eq. 1.

$$\begin{aligned} \hat{\boldsymbol{\tau}}(\hat{Q}; \mathbf{x}) &= \frac{1}{N} \sum_{i=1}^{N} (\hat{Q}(1, \mathbf{x}_i) - \hat{Q}(0, \mathbf{x}_i)), \\ \epsilon_{ATE} &= |\frac{1}{N} \sum_{i=1}^{N} (\hat{\boldsymbol{\tau}}(\hat{Q}; \mathbf{x}_i) - \tau(\mathbf{x}))| \end{aligned} \tag{1}$$

In order to estimate eATE we assume access to the ground truth treatment effect parameter $\boldsymbol{\tau}$, which is only possible with synthetic or semi-synthetic datasets. The Conditional Average Treatment Effect (CATE) may also be calculated on a per-individual basis and the Precision in Estimating Heterogeneous Effect (PEHE) is one way to evaluate a model's efficacy in estimating this quantity:

[2]We use circumflex to designate an estimated (rather than true population) quantity.

[3]For a binary outcome variable $\mathbf{y} \in \{0, 1\}$, $\mathbb{E}(\mathbf{y}|\mathbf{t}, \mathbf{x})$ is the same as the conditional probability distribution $p(\mathbf{y}|\mathbf{t}, \mathbf{x})$.

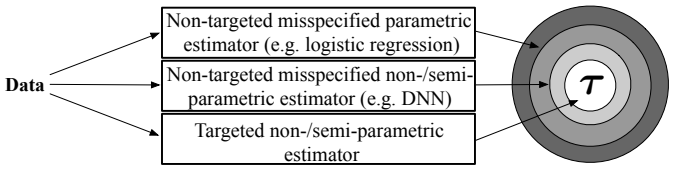

Fig. 2. Different methods for estimating the causal parameter $\boldsymbol{\tau}$ yield different levels of bias. Adapted from [7].

$$\epsilon_{PEHE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\boldsymbol{\tau}}(\hat{Q}; \mathbf{x}_i) - \tau(\mathbf{x}_i))^2} \tag{2}$$

The Naive Approach: The DAG in Fig. 1(a) highlights the problem with taking a naive approach to modeling the joint distribution $p(\mathbf{y}, \mathbf{t}, \mathbf{x})$. The structural relationship $\mathbf{t} \leftarrow \mathbf{x} \rightarrow \mathbf{y}$ indicates both that the assignment of treatment $\mathbf{t}$ is dependent on the covariates $\mathbf{x}$, and that a backdoor path exists through $\mathbf{x}$ to $\mathbf{y}$. In addition to our previous assumptions, if we also assume linearity, adjusting for this backdoor path is a simple matter of adjusting for $\mathbf{x}$ by including it in a logistic regression. The naive method is an example of the uppermost methods depicted in Fig. 2, and leads to the largest bias. The problem with the approach is (a) that the graph is likely misspecified such that the true relationships between covariates as well as the relationships between covariates and the outcome may be more complex. There is also problem (b), that linearity is not sufficient to 'let the data speak' [7] or to avoid biased parameter estimates (*i.e.*, functional misspecification). Using powerful nonparametric models (*e.g.*, neural networks) may solve the limitations associated with linearity and interactions to yield a consistent estimator for $p(\mathbf{y}|\mathbf{X})$, and such a model is an example of the middlemost methods depicted in Fig. 2. However, this estimator is not targeted to the estimation of the causal effect parameter $\boldsymbol{\tau}$, only predicting the outcome, and we require a means to reduce residual bias, such as targeted learning [7]–[9].

Targeted Learning: Targeted Maximum Likelihood Estimation (TMLE) [7]–[9] involves three main steps: (1) estimation of the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{t}, \mathbf{x})$ with estimator $\hat{Q}^0(\mathbf{t}, \mathbf{x})$, (2) estimation of the propensity scores with estimator $\hat{g}(\mathbf{t}|\mathbf{x})$, and (3) updating the conditional mean estimator $\hat{Q}^0$ to get $\hat{Q}^*$ using the propensity scores to attain an estimate for the causal parameter $\boldsymbol{\tau}$.

The propensity score for individual i is defined as the conditional probability of being assigned treatment $g(t_i, \mathbf{x}_i) = p(\mathbf{t} = t_i|\mathbf{x} = \mathbf{x}_i), \in [0, 1]$ [17]. The scores can be used to compensate for the relationship between the covariates and the treatment assignment using Inverse Probability of Treatment Weights (IPTWs), reweighting each sample according to its propensity score. Step (3) is undertaken using 'clever covariates' which are similar to the IPTWs. They form an additional covariate variable $H(1, \mathbf{x}_i) = g(1|\mathbf{x}_i)^{-1}$ for individual i assigned treatment, and $H(0, \mathbf{x}_i) = -g(1|\mathbf{x}_i)^{-1}$ for individual i not assigned treatment. The notation when

conditioning on a single numeric value implies an intervention (*i.e.*, $g(1|\mathbf{x}_i) \equiv g(\text{do}(\mathbf{t} = 1)|\mathbf{x}_i)$). A logistic regression is then undertaken: $\mathbf{y} = \sigma^{-1}[\hat{Q}^0(\mathbf{t}, \mathbf{x})] + \epsilon\hat{H}(\mathbf{t}, \mathbf{x})$ where σ^{-1} is the logit/inverse sigmoid function, $\hat{Q}^0(\mathbf{t}, \mathbf{x})$ is set as a constant, suppressed offset (*i.e.*, no associated regression coefficient) and ϵ represents a fluctuation parameter which is to be estimated from the regression. Once ϵ has been estimated, we acquire an updated estimator:

$$\hat{Q}^1(\text{do}(\mathbf{t} = t), \mathbf{x}) = \sigma\left[\sigma^{-1}[\hat{Q}^0(\mathbf{t}, \mathbf{x})] + \hat{\epsilon}\hat{H}(\mathbf{t}, \mathbf{x})\right] \tag{3}$$

This equation tells us that our new estimator $\hat{Q}^1$ is equal to the old estimator balanced by the corrective $\epsilon H(\mathbf{t}, \mathbf{x})$ term. This term adjusts for the bias associated with the propensity scores. When the ϵ parameter is zero, it means that there is no longer any influence from the 'clever covariates' $H()$. The updated estimator $\hat{Q}^1$ can then be plugged into the estimator for $\hat{\boldsymbol{\tau}}(\hat{Q}^1; \mathbf{x})$. When the optimal solution is reached (*i.e.*, when $\epsilon = 0$), the estimator $\hat{Q}^*$ also satisfies what is known as the efficient Influence Curve (IC), or canonical gradient equation [7], [18], [19]:

$$0 = \sum_{i=1}^{N} IC^*(y_i, t_i, \mathbf{x}_i) = \sum_{i=1}^{N}\Big[\hat{H}(t_i, \mathbf{x}_i)(y_i - \hat{Q}(t_i, \mathbf{x}_i)) + \hat{Q}(1, \mathbf{x}_i) - \hat{Q}(0, \mathbf{x}_i) - \boldsymbol{\tau}(Q; \mathbf{x})\Big] \tag{4}$$

where $IC(y_i, t_i, \mathbf{x}_i)$ represents the IC, and $IC^*(y_i, t_i, \mathbf{x}_i)$ represents the efficient IC for consistent $\hat{Q}$ and $\hat{g}$. It can be seen from the right hand side Eq. 4 that at convergence, the estimator and its corresponding estimand are equal: $y_i = \hat{Q}(t_i, \mathbf{x}_i)$ and $\hat{Q}(1, \mathbf{x}_i) - \hat{Q}(0, \mathbf{x}_i) = \boldsymbol{\tau}(Q; \mathbf{x})$. Over the whole dataset, all terms in Eq. 4 'cancel' resulting in the mean $\bar{IC} = 0$. As such, the logistic regression in Eq. 3 represents a solution to the IC via a parametric submodel.

The TMLE method provides a doubly robust, asymptotically efficient estimate of the causal or 'target' parameter, and these theoretical guarantees make it attractive for adaptation into neural networks for causal effect estimation.

III. Methodology

In this section we present the Targeted Variational AutoEncoder (TVAE), a deep generative latent variable model that enables estimation of the average and conditional average treatment effects (ATE and CATE resp.) via the combination of amortized variational inference techniques and Targeted Maximum Likelihood Estimation (TMLE). A top-level diagram for TVAE is shown in Fig. 3 and follows the structure implied by the DAG in Fig. 1(c).

Assumptions: As is common [15], [20]–[22] when undertaking causal inference with observational data, we make a number of assumptions: (1) Stable Unit Treatment Value Assumption (SUTVA): the potential outcomes for each individual or data unit are independent of the treatments assigned to all other individuals, such that there are no interactions between individuals. (2) Positivity: the assignment of treatment probabilities are all non-zero and non-deterministic $p(\mathbf{t} = t_i|\mathbf{x} = \mathbf{x}_i) > 0, \forall\ \mathbf{t}$ and $\mathbf{x}$. (3) That all confounders have been inferred via noisy proxies present in the observed data [23], [24], such that the likelihood of treatment for two individuals with the same inferred latent covariates is equal, and the potential outcomes for two individuals with the same latent covariates are also equal $s.t.\ \mathbf{y}(1), \mathbf{y}(0) \perp\!\!\!\perp \mathbf{t}|\mathbf{z}$ and $\mathbf{t} \perp\!\!\!\perp (\mathbf{y}(1), \mathbf{y}(0))|\mathbf{z}$, which constitutes a form of conditional exchangeability. In relation to assumption (3), see discussion below on identifiability.

TVAE: If one had knowledge of the true causal DAG underlying a set of data, one could undertake causal inference without being concerned for issues relating to structural misspecification. Unfortunately, and this is particularly the case with observational data, we rarely have access to this knowledge. Quite often an observed set of covariates $\mathbf{x}$ are modelled as a group of confounding variables (as per the DAG in Figure 1a). Furthermore, and as noted by [11], researchers may in general be encouraged to incorporate as many covariates into their model as possible, in an attempt to reduce the severity of the ignorability assumption. However, including more covariates than is necessary leads to other problems relating to the curse of dimensionality and (in)efficiency of estimation.

A large set of covariates may be separable into subsets of factors such as instrumental, risk, and confounding factors. Doing so helps us to match our model more closely to the true data generating process, as well as to improve estimation efficiency by 'distilling' our covariate adjustment set. Prior work has explored the potential to discover the relevant confounding covariates via Bayesian networks [25], regularized regression [26], and deep latent variable models based on Variational Autoencoders (VAEs) [11], [24], [27]. The first two methods *identify* variables (and are variable selection algorithms), whereas VAEs *infer* them, and learn compact, disentangled representations of the observations. The benefit of the latter approach is that it (a) infers latent variables on a datapoint-by-datapoint basis (rather than deriving subsets from population aggregates), (b) under additional assumptions, VAEs have been shown to infer hidden confounders in the presence of noisy proxy variables, thereby potentially reducing the reliance on ignorability [24], [28], and (c) makes no assumptions about the functional form used to map between covariate and latent space.

Variational Inference: In general terms, variational inference is concerned with maximising what is known as the Evidence Lower BOund (ELBO) [29], which constitutes a bound on the likelihood of the data. The ELBO can be derived using the relationship in Eq. 5:

$$\log p_\theta(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right] \tag{5}$$

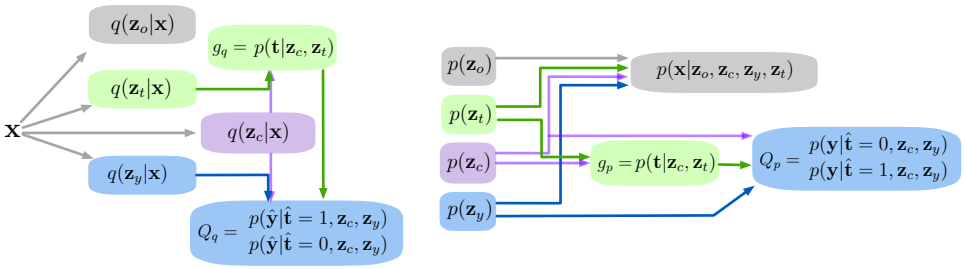

Fig. 3. The block-diagram for Targeted VAE. Dashed boxes indicate the variationally inferred latent variables $\mathbf{z}_o$, $\mathbf{z}_c$, $\mathbf{z}_t$, and $\mathbf{z}_y$. Arrows indicate functions, and colors distinguish treatment (green), outcome (blue), covariates (grey), and confounder (purple) related entities.

Here, the second term on the right hand side represents the Kullback-Liebler divergence between the approximating posterior $q(\mathbf{z}|\mathbf{x})$ and the true posterior $p(\mathbf{z}|\mathbf{x})$. As we do not have access to the true posterior, we specify a prior $p(\mathbf{z})$ as well as a family of (tractable) posterior distributions, and minimize the divergence between the prior and the posterior whilst also attempting to maximise the model evidence. The VAE provides the means to scale this amortized variational inference to intractable, high-dimensional problems, and minimizes the negative log likelihood over a dataset of N samples by adjusting the parameters of neural networks $\{\theta, \phi\}$ according to the ELBO:

$$\frac{1}{N}\sum_{i=1}^{n} -\log p_\theta(\mathbf{x}_i) \leq$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log p_\theta\left(\mathbf{x}_i|\mathbf{z}\right)\right] + \beta\mathbb{D}_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}_i)\|p(\mathbf{z})\right]\right) \tag{6}$$

where $\beta = 1$ is used for the standard variational approximation procedure, but may be set empirically [30], annealed [31] or optimized according to the Information Bottleneck principle [32], [33]. The first term in Eq. 6 is the negative log-likelihood and is calculated in the form of a reconstruction error. The second term is the KLD between the approximating posterior and the prior, and therefore acts as a prior regularizer. Typically, the family of isotropic Gaussian distributions is chosen for the posterior $q_\phi(.)$, and an isotropic Gaussian with unit variance for the prior $p(\mathbf{z})$.

Structure: We seek to infer and disentangle the latent distribution into subsets of latent factors using VAEs. These latent subsets are $\{\mathbf{z}_t, \mathbf{z}_y, \mathbf{z}_c, \mathbf{z}_o\}$, which represent the instrumental factors on $\mathbf{t}$, the risk factors on $\mathbf{y}$, the confounders on both $\mathbf{t}$ and $\mathbf{y}$, and factors solely related to $\mathbf{x}$, respectively. Without inductive bias, consistently disentangling the latent variables into these factors would be impossible [34] because there would be no guidance with which to assign specific information into specific latent factors. In TVAE this inductive bias is incorporated in a number of ways: firstly, by incorporating supervision and constraining $\mathbf{z}_t$ and $\mathbf{z}_y$ to be predictive of $\mathbf{t}$ and $\mathbf{y}$, respectively; secondly, by constraining $\mathbf{z}_c$ to be predictive of both $\mathbf{t}$ and $\mathbf{y}$; and finally, by employing diagonal-covariance priors (isotropic Gaussians) to encourage disentanglement and independence between latent variables. The structural inductive bias on the model is such that $\mathbf{z}_y$, and $\mathbf{z}_t$, and $\mathbf{z}_c$ learn factors relevant to outcome and treatment, for which we provide explicit supervision, thereby leaving $\mathbf{z}_o$ for all remaining factors.

Identifiability: In general, it is impossible to isolate the effect of $\mathbf{t} \rightarrow \mathbf{y}$ due to unobserved confounding [35], and this is why we make the assumption that the latent parents of the treatment and outcome may be inferred via noisy proxies present in the observed data. Under such assumptions, deep latent variable techniques have been shown to be able to infer hidden confounders from these proxy variables (see *e.g.*, [23], [24], [28], [36], [37]). The usual assumption of ignorability then shifts from 'all confounders are observed', to 'all unobserved confounders have been inferred from proxies', both of which represent conditional exchangeability: $y(t) \perp\!\!\!\perp \mathbf{t}|\mathbf{z}$, such that the potential outcome is independent of the observed treatment given the inferred latent variables. We note that, empirically, variational models are not immune to convergence difficulties during optimization which may affect identifiability (see *e.g.* [38] for an exploration of the limitations of VAEs in the context of causal inference), and further work is required to establish bounds on the efficacy of these methods.

The proof for identifiability under the assumption of ignorability on the basis that relevant parent factors have been inferred from proxies and/or other observed variables, has been derived previously by [24] and [11]. With reference to the graph depicted in Fig. 1(c), the factor $\mathbf{z}_o$ is d-separated from $\mathbf{t}$ and $\mathbf{y}$ given $\mathbf{z}$ (or $\mathbf{x}$), and does not affect the identification of the causal effect. *i.e.*, the outcome under intervention $p(y|\text{do}(t), \mathbf{x})$ can be estimated from observational quantities given the inferred latent instrumental, risk, and confounding variables: $p(y|\text{do}(t), \mathbf{x}) = p(y|\text{do}(t), \mathbf{z}_{\{t,o,y,c\}})$. In turn, following the Markov property, $p(y|\text{do}(t), \mathbf{z}_{\{t,o,y,c\}}) = p(y|t, \mathbf{z}_y, \mathbf{z}_c)$ (see [11] and [24] for additional information).

Implementation: We impose the priors and parameterizations denoted in Equations 7 and 8 , where $D_{(.)}$ is the number of dimensions in the respective variable (latent or otherwise), and f_{1-11} and h_{1-6} represent fully connected neural network functions. Note that all Gaussian variance parameterizations are diagonal. In cases where prior knowledge dictates a discrete rather than continuous outcome, equivalent

parameterizations to those in Eqs. 7 and 8 may be employed. For example, in the IHDP dataset, the outcome data are standardized to have a variance of 1, and the outcome generation model becomes a Gaussian with variance also equal to 1. Note that separate treatment and outcome classifiers are used both during inference and generation ($\hat{Q}_q, \hat{g}_q$ and $\hat{Q}_p, \hat{g}_p$ resp.). The classifiers for inference have separate parameters to those use during generation. Predictors or classifiers of outcome incorporate the two-headed approach of [39], and ground-truth $\mathbf{t}$ are used for $\hat{Q}_q$ whereas generated samples $\hat{\mathbf{t}}$ are used for $\hat{Q}_p$. For unseen test cases, either the ground-truth $\mathbf{t}$ or an sampled treatment $\hat{\mathbf{t}}$ from treatment classifier $\hat{g}_p$ may be used to simulate an outcome. During training $\hat{\mathbf{t}}$ is used.

Inference:

$$
\begin{aligned}
q(\mathbf{z}_t|\mathbf{x}) &= \prod_{d=1}^{D_{z_t}} \mathcal{N}(\mu_d = f_{1d}(\mathbf{x}), \sigma_d^2 = f_{2d}(\mathbf{x})) \\
q(\mathbf{z}_y|\mathbf{x}) &= \prod_{d=1}^{D_{z_y}} \mathcal{N}(\mu_d = f_{3d}(\mathbf{x}), \sigma_d^2 = f_{4d}(\mathbf{x})) \\
q(\mathbf{z}_c|\mathbf{x}) &= \prod_{d=1}^{D_{z_c}} \mathcal{N}(\mu_d = f_{5d}(\mathbf{x}), \sigma_d^2 = f_{6d}(\mathbf{x})) \\
q(\mathbf{z}_o|\mathbf{x}) &= \prod_{d=1}^{D_{z_o}} \mathcal{N}(\mu_d = f_{7d}(\mathbf{x}), \sigma_d^2 = f_{8d}(\mathbf{x})) \\
p(\hat{\mathbf{t}}|\mathbf{z}_t, \mathbf{z}_c) &= \text{Bern}(\hat{g}_q(.)) = \text{Bern}(f_9(\mathbf{z}_t, \mathbf{z}_c)) \\
p(\hat{\mathbf{y}}|\mathbf{z}_y, \mathbf{z}_c, \mathbf{t}) &= \text{Bern}(\hat{Q}_q(.)) = \\
&\text{Bern}(\hat{\mathbf{t}} \cdot f_{10}(\mathbf{z}_y, \mathbf{z}_c) + (1 - \hat{\mathbf{t}}) \cdot f_{11}(\mathbf{z}_y, \mathbf{z}_c))
\end{aligned}
\tag{7}
$$

Generation:

$$
\begin{aligned}
p(\mathbf{z}_{\{o,t,c,y\}}) &= \prod_{d}^{D_{\{z_{o,t,c,y}\}}} \mathcal{N}(z_{\{o,t,c,y\}d}|0,1) \\
p(\hat{\mathbf{t}}|\mathbf{z}_t, \mathbf{z}_c) &= \text{Bern}(g_p(.)) = \text{Bern}(h_1(\mathbf{z}_t, \mathbf{z}_c)) \\
p(\hat{\mathbf{y}}|\mathbf{z}_y, \mathbf{z}_c, \hat{\mathbf{t}}) &= \text{Bern}(\hat{Q}_p(.)) = \\
&\text{Bern}(\hat{\mathbf{t}} \cdot h_2(\mathbf{z}_y, \mathbf{z}_c) + (1 - \hat{\mathbf{t}}) \cdot h_3(\mathbf{z}_y, \mathbf{z}_c)) \\
p(\hat{\mathbf{x}}_{bin}|\mathbf{z}_c, \mathbf{z}_o, \mathbf{z}_t, \mathbf{z}_y) &= \text{Bern}(h_6(\mathbf{z}_c, \mathbf{z}_o, \mathbf{z}_t, \mathbf{z}_y)) \\
p(\hat{\mathbf{x}}_{cont}|\mathbf{z}_c, \mathbf{z}_o, \mathbf{z}_t, \mathbf{z}_y) &= \\
\prod_{d=1}^{D_{x_{cont}}} \mathcal{N}(x_{cont,d}|\mu_d = h_4(\mathbf{z}_{\{c,o,t,y\}}), \sigma_d^2 &= h_5(\mathbf{z}_{\{c,o,t,y\}}))
\end{aligned}
\tag{8}
$$

The parameters for these neural networks are learnt via variational Bayesian approximate inference [40] according to the objective in Eq. 9:

$$
\begin{aligned}
\mathcal{L}_i^{\text{ELBO}} = \sum_i^N \mathbb{E}_{q_c q_t q_y q_o} &\left[\log p\left(\hat{\mathbf{x}}_i|\mathbf{z}_t, \mathbf{z}_c, \mathbf{z}_y, \mathbf{z}_o\right)\right. \\
&\left. + \log p\left(\hat{t}_i|\mathbf{z}_t, \mathbf{z}_c\right) + \log p\left(\hat{y}_i|t_i, \mathbf{z}_y, \mathbf{z}_c\right)\right] \\
- \left[D_{KL}\left(q\left(\mathbf{z}_t|\mathbf{x}_i\right)||p\left(\mathbf{z}_t\right)\right) + D_{KL}\left(q\left(\mathbf{z}_c|\mathbf{x},\right)\|p\left(\mathbf{z}_c\right)\right)\right. \\
\left. + D_{KL}\left(q\left(\mathbf{z}_y|\mathbf{x}_i\right)\|p\left(\mathbf{z}_y\right)\right) + D_{KL}\left(q\left(\mathbf{z}_o|\mathbf{x}_i\right)\|p\left(\mathbf{z}_o\right)\right)\right]
\end{aligned}
\tag{9}
$$

Targeted Regularization: We now introduce the targeted regularization, the purpose of which is to encourage the outcome to be independent of the treatment assignment. Following Eq. 3, we define the fluctuation sub-model and corresponding logistic loss for estimating ϵ in Eqs. 10 and 11:

$$
\begin{aligned}
\hat{Q}^1(\hat{g}, t_i, \mathbf{z}_i^y, \mathbf{z}_i^c, \hat{\epsilon}) = \sigma &\left[\sigma^{-1}[\hat{Q}^0(t_i, \mathbf{z}_i^y, \mathbf{z}_i^c)]\right. \\
&\left. + \hat{\epsilon}\left(\frac{I(t_i = 1)}{\hat{g}(t_i = 1; \mathbf{z}_i^t, \mathbf{z}_i^c)} - \frac{I(t_i = 0)}{\hat{g}(t_i = 0; \mathbf{z}_i^t, \mathbf{z}_i^c)}\right)\right]
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
\xi_i(\hat{Q}^1; \hat{\epsilon}) = &-y_i \log(\hat{Q}^1(\hat{g}, t_i, \mathbf{z}_i^y, \mathbf{z}_i^c, \hat{\epsilon})) \\
&-(1 - y_i) \log(1 - \hat{Q}^1(\hat{g}, t_i, \mathbf{z}_i^y, \mathbf{z}_i^c, \hat{\epsilon}))
\end{aligned}
\tag{11}
$$

In Eq. 10, I is the indicator function. For an unbounded regression loss, mean squared error loss may be used. Note that the logistic loss is suitable for continuous outcomes bounded between 0 and 1 (see [7, pp.121:132] for proof). Putting it all together, we then optimize to find generative parameters for functions h_{1-6}, inference parameters for functions f_{1-12}, and estimated fluctuation parameter $\hat{\epsilon}$ in Eq. 12:

$$
\begin{aligned}
\mathcal{L} = \min\left[\sum_i^N\left(\mathcal{L}_i^{\text{ELBO}} + \lambda_{TL}\xi_i(\hat{Q}, \hat{g}, \hat{\epsilon})\right)\right]; \\
\frac{\partial}{\partial \epsilon}\mathcal{L}^*\Big|_{\epsilon=0} = \bar{IC}^* = 0
\end{aligned}
\tag{12}
$$

In Eq. 12, λ_{TL} represents a hyperparameter for the targeted regularization weight. At convergence $\hat{\epsilon} = 0$ and $\hat{Q}$ and $\hat{g}$ become consistent estimators, thereby satisfying the conditions for the EIC (see Eq. 4 and reference [7, pp125:128]).

One aspect of TVAE that bears mentioning (and which differentiates TVAE from another recent contribution [41] that uses targeted regularization) is that the gradients resulting from ξ are *not* taken with respect to the propensity score arms $\hat{g}_p$ or $\hat{g}_q$. Targeted learning is concerned with debiasing the outcome classifier $\hat{Q}$ using propensity scores from $\hat{g}$. In other words, assuming the propensity scores are consistently estimated, the targeted learning regularizer is intended to affect the outcome classifier only, and *not* the propensity score estimator. It is therefore more theoretically aligned (with the targeted learning literature) to apply regularization to the outcome estimator $\hat{Q}$, and not to $\hat{g}$. As per Eq. 3, in TMLE, $\hat{g}$ is assumed to be a consistent estimator, forming part of the debiasing update process for $\hat{Q}$, but it is not subject to update itself. In order to prevent the regularization from affecting the propensity arms, the gradients from the regularizer are only taken with respect to all parameters that influence this outcome classifier (which include upstream parameters for $\hat{Q}_q$ as well as the more direct parameters $\hat{Q}_p$). We use Pytorch's 'detach' method on the propensity scores when calculating the targeted regularization. This method decouples the propensity

score arm from backpropagation relating to the computation of the regularization value.

The notable aspects of our model are as follows: the introduction of a new latent variable $\mathbf{z}_o$ for factors unrelated to outcome and/or treatment to aid the recovery of the true underlying structure and, as far as we are aware, the first incorporation of targeted learning in a deep latent variable approach.

IV. Related Work

There are a number of ways to mitigate the problems associated with the confounding between the covariates and the treatment. For a review on such methods, readers are pointed to the recent surveys by [12], [20]. Here we consider methods that utilize neural networks as part of their models, but note that many non-neural network methods exist [7], [22], [42]–[44].

Perhaps the most similar works to ours are those of Dragonnet [41], TEDVAE [11], and Intact-VAE [27]. We discuss the differences between these and TVAE in turn. Dragonnet is a non-generative model which incorporates the same targeted learning regularization process which allows for the simultaneous optimization of Q and ϵ. However, the method sacrifices the ability to estimate conditional treatment effects to achieve good estimation of the average treatment effect across the sample. Indeed, they do not report PEHE. Finally, Dragonnet applies regularization to the entire network, whereas we 'target' the regularization to the outcome prediction arm by restricting the backpropagation of gradients.

TEDVAE is one of the few variational generative models for causal inference which builds on CEVAE [24] and seeks disentanglement of the latent instrumental, risk, and confounding factors. However, it has no means to allocate latent variables that are unrelated to treatment and/or outcome. The advantage of including factors $\mathbf{z}_o$ with a variational penalty is that the model has the option to use them, or not to use them, depending on whether they are necessary (*i.e.*, KL is pushed to zero). It is important not to force factors unrelated to treatment and outcome into $\mathbf{z}_{\{c,y,t\}}$ because doing so restricts the overlap between the class of models that can be represented using TEDVAE, and the class of models describing the true process.

Intact-VAE seeks to address some of the issues associated with the use of deep latent variable models and causal identification, in particular when the method's performance hinges on its ability to recover the posterior distribution of unobserved confounders. They do not disentangle risk and instrumental variables from confounders, focusing instead on the successful recovery of the confounder via the use of a novel development of the prognostic score which they refer to as B^*-scores.

Other methods include GANITE [45] which requires adversarial training, and may therefore be more difficult to optimise. PM [46], SITE [47], and MultiMBNN [48] incorporate propensity score matching. TARNET [39] inspired the two-headed outcome arm in our TVAE, as well as the three-headed architecture in [41]. RSB [49] incorporates regularization based on the Pearson Correlation, intended to reduce the association between latent variables predictive of treatment assignment and those predictive of outcome.

V. Experiments

We begin by performing an ablation study on our synthetic TVAESynth dataset, comparing (a) TVAE (base) which is equivalent to TEDVAE (b) TVAE with $\mathbf{z}_o$, and (c) TVAE with both with $\mathbf{z}_o$ and targeted regularization ξ during training. In order to evaluate the benefits of introducing $\mathbf{z}_o$, we ensure that the total number of latent dimensions remains constant.

We then utilize 100 replications of the semi-synthetic *Infant Health and Development Program (IHDP)* dataset [43], [50][4] The linked version (footnote) corresponds with usual setting A of the NPCI data generating package [51] (see [39], [41], [47]) and comprises 608 untreated and 139 treated samples (747 in total). There are 25 covariates, 19 of which are discrete/binary, and the rest are continuous. The outcome for the IHDP data is continuous and unbounded. Similarly to [24], [39] and others, we utilize a 60/30/10 train/validation/test split. We evaluate our network on the Average Treatment Effect estimation error (eATE), and the Precision in Estimation of Heterogeneous Effect (PEHE).

We also utilize the job outcomes dataset (*Jobs*) [52], [53].[5] Unlike the IHDP dataset, Jobs is real-world data with a binary outcome. We follow a similar procedure to [39] who indicate that they used the Dehejia-Wahba [54] and PSID comparison sample. This sample comprises 260 treated samples and 185 control samples, along with the PSID comparison group comprising 2490 samples'. The dataset contains a mixture of observational and RCT data. Similarly to [24], [39] and others, we utilize a 56/24/20 train/validation/test split, and undertake 100 runs with varying random split allocations in order to acquire an estimate of average performance and standard error. Note that, between models, the same random seed is used both for initialization as well as dataset splitting, and therefore the variance due to these factors is equivalent across experiments. As per [24], [39], [47], for the Jobs dataset (for which we have only partial effect supervision) we evaluate our network on the Average Treatment effect on the Treated error:

$$\begin{aligned} eATT = ||T_1|^{-1} \sum_{i \in T_1} y_i - |T_0|^{-1} \sum_{j \in T_0} y_j \\ -|T_1|^{-1} \sum_{i \in T_1} (\hat{Q}(1, \mathbf{x}_i) - \hat{Q}(0, \mathbf{x}_i))| \end{aligned} \tag{13}$$

where $T = T_1 \cup T_0$ constitutes all individuals in the RCT, and the subscripts denote whether or not those individuals were in the treatment (subscript 1) or control groups (subscript 0). The first two terms in Eq. 13 comprise the true ATT, and the third term the estimated ATT. We may use the policy risk as a proxy for PEHE:

[4]Available from https://www.fredjo.com/

[5]Available from https://users.nber.org/~rdehejia/data/.nswdata2.html.

$$\begin{aligned}\mathcal{R}_{pol} = 1 - (&\mathbb{E}[\mathbf{y}(1)|\pi(\mathbf{x}) = 1]p(\pi(\mathbf{x}) = 1) \\ &+\mathbb{E}[\mathbf{y}(0)|\pi(\mathbf{x}) = 0]p(\pi(\mathbf{x}) = 0))\end{aligned} \tag{14}$$

where $\pi(\mathbf{x}_i) = 1$ is the policy to treat when $\hat{y}_i(1) - \hat{y}_i(0) > \alpha$, and $\pi(\mathbf{x}_i) = 0$ is the policy not to treat otherwise [39], [47]. α is a treatment threshold. This threshold can be varied to understand how treatment inclusion rates affect the policy risk. We set $\alpha = 0$, as per [24], [39].

Finally, we introduce a new synthetic dataset named *TVAESynth* which follows the structure shown in Figure 4, and relationships in Equations 15-18. While the weightings are chosen relatively arbitrarily, the structure is intentionally designed such that there are a mixture of exogenous and endogenous covariates. This enables us to compare the performance of TVAE with and without $\mathbf{z}_o$ (keeping the total number of latent dimensions constant). The dataset was designed such that not all covariates are exogenous and so that there exist some latent factors unrelated to outcome and treatment. Thus, we should expect an improvement in performance to occur with the introduction of $\mathbf{z}_o$, demonstrating the importance of incorporating inductive bias that closely matches the true structure. Note that while these datasets vary in whether the outcome variable is continuous (IHDP, TVAESynth) or binary (Jobs), the treatment variable is always binary. We leave an application to data with continuous treatment effects to future work.

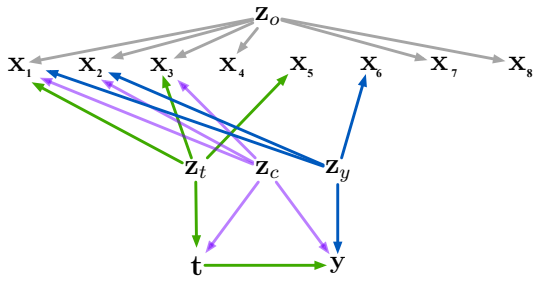

Fig. 4. The DAG for TVAESynth dataset.

$$\begin{aligned}\mathbf{U}_{z_o,z_c,z_t,z_y,y} &\sim \mathcal{N}(\mathbf{0},\mathbf{1}) \\ \mathbf{U}_{x_1,x_4,t} \sim \text{Bernoulli}(0.5) \quad \mathbf{U}_{x_{2:3},x_{5:8}} &\sim \mathcal{N}(\mathbf{0},\mathbf{1}) \\ \mathbf{z}_o = \mathbf{U}_{z_o} \quad \mathbf{z}_y = \mathbf{U}_{z_y} \quad \mathbf{z}_t = \mathbf{U}_{z_t} \quad \mathbf{z}_c &= \mathbf{U}_{z_c}\end{aligned} \tag{15}$$

$$\begin{aligned}\mathbf{x}_1 &\sim \text{Bernoulli}(\sigma(\mathbf{z}_t + 0.1(\mathbf{U}_{x_1} - 0.5))) \\ \mathbf{x}_2 &\sim \mathcal{N}(0.4\mathbf{z}_o + 0.3\mathbf{z}_c + 0.5\mathbf{z}_y + 0.1\mathbf{U}_{x_2}, 0.2) \\ \mathbf{x}_3 &\sim \mathcal{N}(0.2\mathbf{z}_o + 0.2\mathbf{z}_c + 1.2\mathbf{z}_t + 0.1\mathbf{U}_{x_3}, 0.2) \\ \mathbf{x}_4 &\sim \text{Bernoulli}(\sigma(0.6\mathbf{z}_o + 0.1(\mathbf{U}_{x_4} - 0.5))) \\ \mathbf{x}_5 &\sim \mathcal{N}(0.6\mathbf{z}_t + 0.1\mathbf{U}_{x_5}, 0.1) \\ \mathbf{x}_6 &\sim \mathcal{N}(0.9\mathbf{z}_y + 0.1\mathbf{U}_{x_6}, 0.1) \\ \mathbf{x}_7 &\sim \mathcal{N}(0.5\mathbf{z}_o + 0.1\mathbf{U}_{x_7}, 0.1) \\ \mathbf{x}_8 &\sim \mathcal{N}(0.5\mathbf{z}_o + 0.1\mathbf{U}_{x_8}, 0.1)\end{aligned} \tag{16}$$

$$\mathbf{t}_p = \sigma(0.2\mathbf{z}_c + 0.8\mathbf{z}_t + 0.1\mathbf{U}_t) \quad \mathbf{t} \sim \text{Bernoulli}(\mathbf{t}_p) \tag{17}$$

$$y := 0.2\mathbf{z}_c + 0.5\mathbf{z}_y\mathbf{t} + 0.2\mathbf{t} + 0.1\mathbf{U}_\mathbf{y} \tag{18}$$

When estimating treatment effects, 100 samples are drawn for each set of input covariates. We provide results for both within sample and out-of-sample performance. Note that within sample and out-of-sample results are equally valid for treatment effect estimation because the network is never supervised on treatment effect [41].

A. Architecture and Hyperparameters

The architecture is shown in Fig. 5. For continuous outcomes we standardize the values and model as a Gaussian with a fixed variance of 1, and a mean determined by the outcome arm. All binary variables in the model are modelled as Bernoulli distributed with a probability determined by the associated neural network function.

Hyperparameters[6] for IHDP experiments were: hidden layers: **3**; the weight on targeted regularization $\lambda_{TL} = \{0.0, 0.1, 0.2, \mathbf{0.4}, 0.6, 0.8, 1.0\}$; learning rate $LR = \{$1e-3, 1e-4, **5e-5**$\}$; hidden neurons =**300**; layers = **4**; dimensionality of latent factors was $D_{z_t} = D_{z_t} = \mathbf{10}, D_{z_c} = \mathbf{15}, D_{z_o} = \mathbf{5}$; batch size of **200**; epochs **200**; weight regularization **1e-4**; and learning rate decay **5e-4**. Hyperparameters for Jobs experiments were: hidden layers: **3**; the weight on targeted regularization $\lambda_{TL} = \{0.0,$ **0.1**, 0.2, 0.4, 0.6, 0.8, $1.0\}$; learning rate $LR = \{$5e-5, **1e-5**$\}$; hidden neurons =**200**; layers = **2**; dimensionality of latent factors was $D_{z_t} = D_{z_t} = \mathbf{6}, D_{z_c} = \mathbf{8}, D_{z_o} = \mathbf{4}$; batch size of **200**; epochs **200**; weight regularization **1e-4**; and learning rate decay **5e-4=3**. Hyperparameters for TVAESynth experiments were: hidden layers: **2**; the weight on targeted regularization $\lambda_{TL} = \{0.0,$ **0.1**, 0.2, 0.4, 0.6, 0.8, $1.0\}$; learning rate $LR =$ **5e-5**; hidden neurons =**20**; layers = **2**; dimensionality of latent factors was $D_{z_t} = D_{z_t} = D_{z_c} = \mathbf{2}, D_{z_o} = \mathbf{1}$; batch size of **200**; epochs **40**; weight regularization **1e-4**; and learning rate decay **5e-3**. All models were optimized using Adam [55].

For model selection we use the minimum validation loss on the total objective function [11], [24]. Whilst some model selection heuristics exist that serve as surrogates for the eATE itself (*e.g.*, see [44], [56]) we take the same view as [11], in that the development of our model 'should be self-sufficient and not rely on others'. For all experiments, we undertake 100 replications and provide mean and standard error. There may be room to improve performance with further tuning. However, given that the tuning of hyperparameters in a causal inference paradigm is problematic in general, we intentionally limited the search space.

The network is coded using Pyro [57] and is an extension of the code by [11] available here: https://github.com/WeijiaZhang24/TEDVAE. We train on a GPU (*e.g.*, NVIDIA 2080Ti) driven by a 3.6GHz Intel I9-9900K CPU running

[6]Bold font indicates the settings that were finally used.

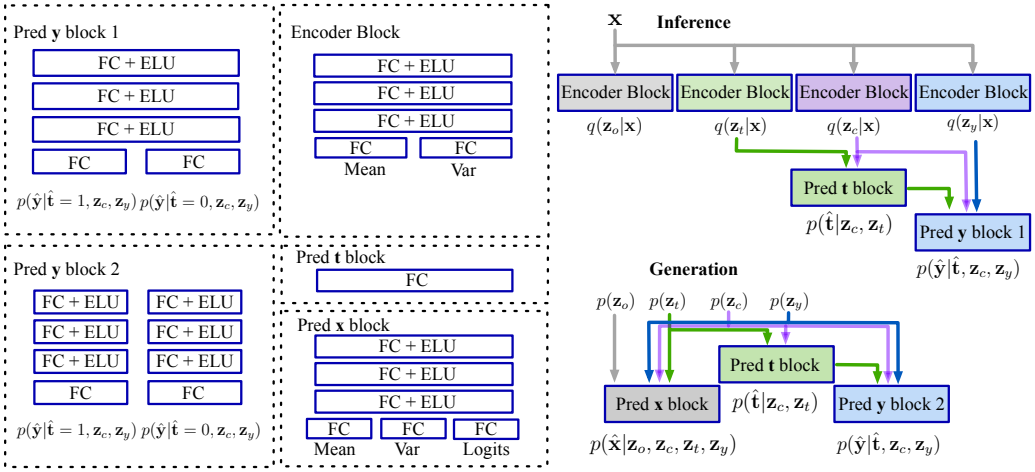

Fig. 5. Block architectural diagram for TVAE's inference and generation models (number of layers in Pred $\mathbf{y}$ 1 & 2 and Pred $\mathbf{x}$ blocks varies by experiment, as does the number of neurons in each layer (see details in text).

Ubuntu 18.04. Training 200 epochs of the IHDP dataset (training split of 450 samples) takes approx. 35 seconds (0.175s per epoch).

B. Results

Ablation Study Results are shown in Table I. They demonstrate that both eATE and PEHE are significantly improved by the incorporation of $\mathbf{z}_o$ or targeted regularization, with a combination of the two yielding the best results for both within sample and out of sample testing. The fact that TVAE $+\mathbf{z}_o$ outperforms TVAE $+\mathbf{z}_o^*$ despite the latter having a larger latent capacity, suggests that reducing the capacity of the latent space has a beneficial, regularizing effect. Based on the results of this ablation, the benefits of this regularizing effect appear to be distinct from the benefits that derive from the addition of miscellaneous factors. Finally, the results indicate negligible empirical benefits to restricting the backgpropagation of the regularizer to non-propensity related parameters. However, our restriction of the backpropagation more closely aligns with the original TMLE and efficient influence curve theory, and we therefore retain this feature for the remaining experiments.

IHDP Results are shown in Table II and indicate state of the art performance for both within sample and out-of-sample eATE and PEHE. They corroborate the ablation results in Table I, in that the incorporation of $\mathbf{z}_o$ and targeted regularization result in monotonic improvement above TED-VAE. TVAE is outperformed only by Dragonnet on within-sample eATE performance. However, this method does not provide estimations for individual CATE, and is limited to the estimation of average treatment effects.

Jobs Results are shown in Table III. GANITE was found to perform the best across most metrics, although this method has been argued to be more reliant on larger sample sizes than others, on the basis that it performs relatively poorly on the smaller IHDP dataset [45]. Furthermore, GANITE relies on potentially unstable/unreliable adversarial training [58]–[60]. Finally, TVAE outperforms GANITE on eATT, is consistent (beyond 2 d.p.) across out-of-sample and within-sample evaluations and has a lower standard err, and is competitive across all metrics. On this dataset, the concomitant improvements associated with the additional latent factors and targeted learning were negligible.

VI. Conclusion

We aimed to improve existing latent variable models for causal parameter estimation in two ways: Firstly, by introducing a latent variable to model factors unrelated to treatment and outcome, thereby enabling the model to more closely reflect the data structure; and secondly, by incorporating a targeted learning regularizer with selected backpropagation to further debias outcome predictions. Our experiments demonstrated concomitant improvements in performance, and our comparison against other methods demonstrated TVAE's ability to compete with and/or exceed state of the art for both individual as well as average treatment effect estimation. In spite of TVAE's promising performance, it is worth remembering the method's limitations. Firstly, we assume that there are sufficient proxy variables present in the observations to facilitate inference of the latent factors. Secondly, variational approaches are approximate and their performance depends on a number of aspects, such as the choice of posterior distribution, and on optimization convergence. The latter point can significantly affect identifiability of causal effects [38].

For future work, we plan apply TVAE to longitudinal data with continuous or categorical treatment, to explore the use of TVAE in inferring hidden confounders from proxies in the dataset, and to explore the bounds on identifiability associated with the use of structured models in combination with targeted regularization. Additionally, it was noted from the ablation study results that the restriction of regularization gradients did not yield a significant change in performance when compared with applying the regularization to the entire network. It would also be highly valuable to establish to what extent the properties of targeted approaches (*e.g.*, double robustness)

TABLE I

MEANS AND STANDARD ERRORS FOR THE ABLATION STUDY USING TVAESYNTH. 'OOS' IS OUT-OF-SAMPLE, 'WS' IS WITHIN SAMPLE. '$+\mathbf{z}_o$' INDICATES THE INTRODUCTION OF THE MISCELLANEOUS FACTORS, '$+\mathbf{z}_o^*$' INDICATES THE INTRODUCTION OF MISCELLANEOUS FACTORS BUT *without* CHANGING THE DIMENSIONALITY OF $\mathbf{z}_c$ THEREBY INCREASING TOTAL LATENT CAPACITY, '$+\xi$' INDICATES OUR TARGETED REGULARIZATION (WITH SELECTED BACKPROPAGATION), '$+\xi^*$' INDICATES TARGETED REG. EQUIVALENT TO THE ONE USED BY [41] (*i.e.*, WITH GRADIENTS APPLIED TO ALL UPSTREAM PARAMETERS), THE ξ SUBSCRIPT INDICATES ITS WEIGHT IN THE LOSS.

Method	$\sqrt{\epsilon_{PEHE}}$ WS	$\sqrt{\epsilon_{PEHE}}$ OOS	ϵ_{ATE} WS	ϵ_{ATE} OOS
TVAE (base/TEDVAE)	.179±.003	.178±.003	.128±.005	.128±.005
TVAE + $\mathbf{z}_o^*$	.174±.003	.173±.003	.121±.005	.120±.005
TVAE + $\mathbf{z}_o$	.166±.003	.166±.003	.069±.004	.069±.004
TVAE + $\xi_{\lambda=0.1}$	.171±.003	.170±.003	.122±.004	.121±.004
TVAE + $\mathbf{z}_o$ + $\xi^*_{\lambda=0.1}$	**.151±.002**	**.150±.003**	**.048±.003**	**.048±.003**
TVAE + $\mathbf{z}_o$ + $\xi_{\lambda=0.1}$ (full model)	**.150±.002**	**.150±.003**	**.048±.003**	**.048±.003**

TABLE II

MEANS AND STANDARD ERRORS FOR EVALUATION ON IHDP [43]. RESULTS FROM: [11], [24], [39], [45]. 'OOS' IS OUT-OF-SAMPLE AND 'WS' IS WITHIN SAMPLE. '$+\mathbf{z}_o$' INDICATES THE INTRODUCTION OF THE MISCELLANEOUS FACTORS, AND '$+\xi$' INDICATES TARGETED REG. WITH SUBSCRIPT INDICATING ITS WEIGHT IN THE LOSS.

Method	$\sqrt{\epsilon_{PEHE}}$ WS	$\sqrt{\epsilon_{PEHE}}$ OOS	ϵ_{ATE} WS	ϵ_{ATE} OOS
TMLE [8]	5.0±.20	-	.30±.01	-
CEVAE [24]	2.7±.10	2.6±.10	.34±.01	.46±.02
TARNet [39]	.88±.00	.95±.00	.26±.01	.28±.01
CFR-MMD [39]	.73±.00	.78±.00	.30±.01	.31±.01
CFR-Wass [39]	.71±.00	.76±.00	.25±.01	.27±.01
TEDVAE [11]	.62±.11	.63±.12	-	.20±.05
IntactVAE [27]	.97±.04	1.0±.05	.17±.01	.21±.01
GANITE [45]	1.9±.40	2.4±.40	.43±.05	.49±.05
Dragonnet w/ t-reg [41]	-	-	**.14±.01**	.20±.01
TVAE ($\mathbf{z}_0$, $\xi_{\lambda=0.0}$)	**.57±.03**	**.57±.03**	**.16±.01**	**.16±.01**
TVAE ($\mathbf{z}_0$, $\xi_{\lambda=0.4}$)	**.52±.02**	**.54±.02**	**.15±.01**	**.16±.01**

TABLE III

MEANS AND STANDARD ERRORS FOR EVALUATION ON JOBS. RESULTS TAKEN FROM: [11], [24]. 'OOS' IS OUT-OF-SAMPLE AND 'WS' IS SAMPLE. '$+\mathbf{z}_o$' INDICATES THE INTRODUCTION OF THE MISCELLANEOUS FACTORS, AND '$+\xi$' INDICATES TARGETED REG. WITH SUBSCRIPT INDICATING ITS WEIGHT IN THE LOSS.

Method	R_{pol} WS	R_{pol} OOS	ϵ_{ATT} WS	ϵ_{ATT} OOS
TMLE [8]	.22±.00	-	.02±.01	-
CEVAE [24]	.15±.00	.26±.00	.02±.01	.03±.01
TARNet [39]	.17±.00	.21±.00	.05±.02	.11±.04
CFR-MMD [39]	.18±.00	.21±.00	.04±.01	.08±.03
CFR-Wass [39]	.17±.00	.21±.00	.04±.01	.09±.03
GANITE [45]	**.13±.00**	**.14±.00**	**.01±.01**	.06±.03
TEDVAE [11]	-	-	.06±.00	.06±.00
TVAE ($\mathbf{z}_0$, $\xi_{\lambda=0}$)	.16±.00	.16±.00	**.01±.00**	**.01±.00**
TVAE ($\mathbf{z}_0$, $\xi_{\lambda=1}$)	.16±.00	.16±.00	**.01±.00**	**.01±.00**

carry over to neural network estimators which use targeted regularization. Finally, the ablation results indicated that part of the improvement associated with the introduction of $\mathbf{z}_o$ is associated with a regularizing effect relating to the reduction in the dimensionality of $\mathbf{z}_c$. This aspect of the model's behavior also deserves investigating.

REFERENCES

[1] N. Kreif and K. DiazOrdaz, "Machine learning in policy evaluation: new tools for causal inference," *arXiv:1903.00402v1*, 2019.

[2] L. Bottou, J. Peters, Q. J., D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: the example of computational advertising," *Journal of Machine Learning Research*, vol. 14, 2013.

[3] M. Petersen, L. Balzer, D. Kwarsiima, N. Sang, G. Chamie, J. Ayieko, J. Kabami, A. Owaraganise, T. Liegler, F. Mwangwa, and K. Kadede, "Association of implementation of a universal testing and treatment intervention with HIV diagnosis, receipt of antiretroviral therapy, and viral suppression in East Africa," *Journal of American Medical Association*, vol. 317, no. 21, pp. 2196–2206, 2017.

[4] J. Pearl, *Causality*. Cambridge: Cambridge University Press, 2009.

[5] B. Siegerink, W. den Hollander, M. Zeegers, and R. Middelburg, "Causal inference in law: an epidemiological perspective," *European Journal of Risk Regulation*, vol. 7, no. 1, pp. 175–186, 2016.

[6] A. Deaton and N. Cartwright, "Understanding and misundertstanding randomized controlled trials," *Social Science and Medicine*, vol. 210, pp. 2–21, 2018.

[7] M. J. van der Laan and S. Rose, *Targeted Learning - Causal Inference for Observational and Experimental Data*. New York: Springer International, 2011.

[8] ——, *Targeted Learning in Data Science*. Switzerland: Springer International, 2018.

[9] M. S. Schuler and S. Rose, "Targeted maximum likelihood estimation for causal inference in observational studies," *American Journal of Epidemiology*, vol. 185, no. 1, pp. 65–73, 2016.

[10] M. J. van der Laan and R. J. C. M. Starmans, "Entering the era of data science: targeted learning and the integration of statistics and computational data analysis," *Advances in Statistics*, 2014.

[11] W. Zhang, L. Liu, and J. Li, "Treatment effect estimation with disentangled latent factors," *arXiv:2001.10652*, 2020.

[12] R. Guo, L. Cheng, J. Li, P. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 1, no. 1, 2020.

[13] J. Pearl, M. Glymour, and N. Jewell, *Causal inference in statistics: A primer*. Wiley, 2016.

[14] M. Vowels, N. Camgoz, and R. Bowden, "D'ya like DAGs? A survey on structure learning and causal discovery," *arXiv:2103.02582*, 2021.

[15] G. Imbens and D. Rubin, *Causal inference for statistics, social, and biomedical sciences. An Introduction.* New York: Cambridge University Press, 2015.

[16] A. Jesson, S. Mindermann, U. Shalit, and Y. Gal, "Identifying causal effect inference failure with uncertainty-aware models," *arXiv:2007.00163v1*, 2020.

[17] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[18] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383–393, 1974.

[19] E. H. Kennedy, *Statistical causal inferences and their applications in public health research*. Spinger, 2016, ch. Semiparametric theory and empirical processes in causal inference.

[20] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *arXiv:2002.02770*, 2020.

[21] R. Guo, J. Li, and H. Liu, "Learning individual causal effects from networked observational data," *Association for Computing Machinery*, 2020.

[22] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[23] M. Montgomery, M. Gragnolati, K. Burke, and E. Paredes, "Measuring living standards with proxy variables," *Demography*, vol. 37, no. 2, pp. 155–174, 2000.

[24] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *31st Conference on Neural Information Processing Systems*, 2017.

[25] J. Haggstrom, "Data-driven confounder selection via Markov and Bayesian networks," *Biometriks*, vol. 74, no. 2, pp. 389–398, 2017.

[26] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, and F. Wang, "Treatment effect estimation with data-driven variable decomposition," *AAAI*, 2017.

[27] P. Wu and K. Fukumizu, "Intact-VAE: Estimating treatment effects under unobserved confounding," *arXiv:2101.06662v2*, 2021.

[28] S. Lowe, D. Madras, R. Zemel, and M. Welling, "Amortized causal discovery: Learning to infer causal graphs from time-series data," *arXiv:2006.10833v1*, 2020.

[29] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: a review for statisticians," *arXiv:1601.00670v9*, 2018.

[30] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," *ICLR*, 2017.

[31] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in Beta-VAE," *arXiv:1804.03599v1*, 2018.

[32] A. A. Alemi, I. FIscher, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv:1612.00410v7*, 2017.

[33] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *arXiv:1503.02406v1*, 2015.

[34] F. Locatello, S. Bauer, M. Lucic, G. Ratsch, S. Gelly, B. Scholkopf, and B. O., "Challenging common assumptions in the unsupervised learning of disentangled representations," *arXiv:1811.12359v3*, 2019.

[35] A. D'Amour, "On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility and alternatives," *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89, 2019.

[36] E. Allman, C. Matias, and J. e. a. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3099–3132, 2009.

[37] S. Parbhoo, M. Wieser, and V. Wieczorek, A.and Roth, "Information bottleneck for estimating treatment effects with systematically missing covariates," *Entropy*, vol. 22, no. 4, 2020.

[38] S. Rissanen and P. Marttinen, "A critical look at the identifiability of causal effects with deep latent variable models," *arXiv2102.06648v1*, 2021.

[39] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," *arxiv:1606.03976v5*, 2017.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114v10*, 2014.

[41] C. Shi, D. M. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," *33rd Conference on Neural Information Processing Systems*, 2019.

[42] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "Double/debiased/Neyman machine learning of treatment effects," *American Economic Review*, vol. 5, 2017.

[43] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, 2011.

[44] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. of the National Academy of Sciences of the USA*, vol. 113, no. 27, 2016.

[45] J. Yoon, J. Jordan, and M. van der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," *ICLR*, 2018.

[46] P. Schwab, L. Linhardt, and W. Karlen, "Perfect match: a simple method for learning representations for counterfactual inference with neural networks," *arXiv:1810.00656v5*, 2019.

[47] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[48] A. Sharma, G. Gupta, A. Prasad, R.and Chatterjee, L. Vig, and G. Shroff, "MultiMBNN: matched and balanced causal inference with neural networks," *arXiv:2004.13446v2*, 2020.

[49] Z. Zhang, Q. Lan, Y. Wang, N. Hassanpour, and R. Greiner, "Reducing selectin bias in counterfactual reasoning for individual treatment effects estimation," *33rd Conference on Neural Information Processing Systems*, 2019.

[50] R. T. Gross, *Infant health and development program (IHDP): enhancing the outcomes of low birth weight, premature infants in the United States*. Ann Arbor: MI: Inter-university Confortium for Political and Social Research, 1993.

[51] V. Dorie, "Non-parametrics for causal inference," *https://github.com/vdorie/npci*, 2016.

[52] R. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, pp. 604–620, 1986.

[53] J. A. Smith and P. E. Todd, "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, vol. 125, no. 1, pp. 305–353, 2005.

[54] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and Statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[55] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," *arXiv:1412.6980v9*, 2017.

[56] N. Hassanpour and R. Greiner, "Counterfactual regression with importance sampling weights," *Proceedings of the 28th International Joint Conference on Artificial Intellgence (IJCAI-19)*, 2019.

[57] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep universal probabilistic programming," *J. Mach. Learn. Res.*, vol. 20, 2019.

[58] D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan, "Invariant representations without adversarial training," *NeurIPS*, 2018.

[59] J. Lezama, "Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision," *ICLR*, 2019.

[60] A. Gabbay and Y. Hosen, "Demystifying inter-class disentanglement," *arXiv:1906.11796v2*, 2019.