Variational Auto-Encoder for Causal Inference

Negar Hassanpour [1] **Russell Greiner** [1]

Abstract

This paper provides a generative approach for causal inference in observational studies. Inspired by the semi-supervised Variational Auto-Encoder (VAE), we propose a novel double-stacked M2 architecture with β-VAE components that encourage learning disentangled representations. Our empirical results demonstrate the superiority of the proposed method compared to both state-of-the-art discriminative as well as generative approaches in the literature.

1. Introduction

As one of the main tasks in studying causality (Peters et al., 2017; Guo et al., 2018), the goal of Causal Inference is to figure out **how much** the value of a certain variable would change (*i.e.*, *effect*) had another certain variable (*i.e.*, cause) changed its value. A prominent example is the counterfactual question (Rubin, 1974; Pearl, 2009) "Would this patient have *lived longer* [and by **how much**], had she received an alternative treatment?". Such question is often asked in the context of precision medicine, that attempts to identify which medical procedure $t \in \mathcal{T}$ will benefit a certain patient x the most, in terms of the treatment outcome $y \in \mathbb{R}$.

Missing values is a major challenge in causal inference; since, for each subject i, any real-world dataset can only contain the outcome of the administered treatment (aka *observed* outcome: y_i), but not the outcome(s) of the alternative treatment(s) (aka *counterfactual* outcome(s) – *i.e.*, y_i^t for $t \in \mathcal{T} \backslash \{t_i\}$ denoted by $\neg t_i$. The fact that counterfactual outcomes are **unobservable** (*i.e.*, missing in any training data) makes estimating treatment effects more difficult than the generalization problem in the supervised learning paradigm.

Most of the current causal inference methods can be categorized as *discriminative* approaches – *i.e.*, they only observe and condition on the provided data and make no efforts to figure out the underlying mechanism that actually generated the data. These include the Balancing Neural Network (BNN) (Johansson et al., 2016), CounterFactual Regression Network (CFR-Net) (Shalit et al., 2017), and CFR-Net's extensions – *cf.*, (Hassanpour & Greiner, 2019; 2020; Yao et al., 2018) – as well as Dragon-Net (Shi et al., 2019).

A promising direction is developing *generative* models, using either Generative Adverserial Network (GAN) (Goodfellow et al., 2014) or Variational Auto-Encoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014). Two generative approaches for causal inference in the literature are: (i) GANs for inference of Individualised Treatment Effects (GANITE) (Yoon et al., 2018) and (ii) Causal Effect VAE (CEVAE) (Louizos et al., 2017). However, neither of the two achieve competitive performance compared to the discriminative approaches.

Contribution: In this paper, we provide a VAE-based generative model for causal inference that significantly enhances state-of-the-art on two publicly available benchmarks.

The rest of this document is organized as follows: Section 2 elaborates on the ideas presented in some of the above-mentioned papers and discusses their main contributions and gaps. Section 3 presents our proposed method: a double-stacked M2 VAE (Kingma et al., 2014). Section 4 summarizes the experiments and discusses the performance results of the proposed method compared to the contenders. Section 5 concludes the paper with future directions of this research and summary of contributions of the current work.

2. Related works

CFR-Net Shalit et al. (2017) learned a representation space Φ to reduce selection bias by making $\Pr(\, x \,|\, t=0\,)$ and $\Pr(\, x \,|\, t=1\,)$ as close to each other as possible (see Figure 1), provided that $\Phi(\, x\,)$ retains enough information such that all the $|\mathcal{T}|$ learned regressors $\{h^t\big(\Phi(\cdot)\big),\ \forall t \in \mathcal{T}\}$ can generalize well on the observed outcomes. Their objective function includes $L\big[\, y_i,\, h^{t_i}\big(\Phi(x_i)\big)\big]$ which is the loss of predicting the observed outcome for sample i, weighted by ω_i. These weights are derived via $\omega_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$, where $u = \frac{1}{N}\sum_{i=1}^{N} t_i = \Pr(\, t=1\,)$. This is effectively setting:

$$\omega_i \;=\; \frac{1}{2\,\Pr(\, t_i\,)} \;=\; \frac{1}{2}\,\big[\,1 + \frac{\Pr(\,\neg t_i\,)}{\Pr(\, t_i\,)}\,\big] \qquad (1)$$

[1]Department of Computing Science, University of Alberta, Edmonton, Canada. Correspondence to: Negar Hassanpour <hassanpo@ualberta.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

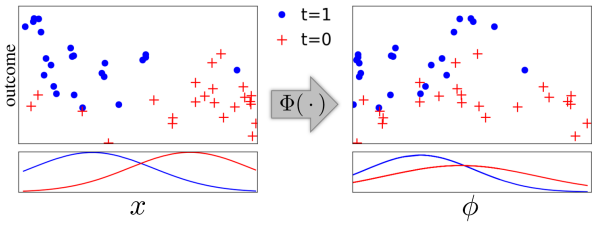

Figure 1. Selection bias can be reduced using representation learning. Note there were few +'s for small values of x (left), but there are +'s across the entire range of ϕ (right); similarly for (•)'s.

where $\Pr(t_i)$ is the probability of selecting treatment $t_i \in \{0, 1\}$ over the entire population.

Dragon-Net Shi et al. (2019)'s main objective was to estimate the ATE, which they explain requires a two stage procedure: (i) fit models that predict the outcomes and (ii) find a downstream estimator of the effect. Their method is based on a classic result from strong ignorability – *i.e.*, Theorem 3 in (Rosenbaum & Rubin, 1983) – which states:

$$(y^1, y^0) \perp\!\!\!\perp t \,|\, x \quad \& \quad \Pr(t = 1 \,|\, x) \in (0, 1) \implies$$
$$(y^1, y^0) \perp\!\!\!\perp t \,|\, b(x) \quad \& \quad \Pr(t = 1 \,|\, b(x)) \in (0, 1)$$

where $b(x)$ is a balancing score[1]. They consider propensity score as a balancing score and argue that only the parts of X relevant for predicting T are required for the estimation of the causal effect.[2] This theorem, however, only provides a way to *match* treated and control instances. In other words, it helps finding potential counterfactual outcomes from the alternative group in order to calculate ATE. Shi et al. (2019), however, used this theorem to derive minimal representations on which to *regress* in order to estimate the outcomes.

GANITE Yoon et al. (2018) proposed the counterfactual GAN, whose generator G, given $\{x, t, y^t\}$, estimates the counterfactual outcomes ($\hat{y}^{\neg t}$); and whose discriminator D tries to identify the factual outcome given $\{x, (y^t, \hat{y}^{\neg t})\}$. It is, however, unclear why G can produce samples that are indistinguishable from the factual outcomes while it is plausible that D can just learn the treatment selection mechanism instead of distinguishing the factual outcomes from counterfactuals. Although this work is among the few generative approaches for causal inference in the literature, our empirical results (*cf.*, Section 4) shows that it is not successful in terms of accurate estimation of counterfactuals.

CEVAE Louizos et al. (2017) used VAE to extract latent confounders from their observed proxies $\{X, T, Y\}$. While this is an interesting step in the right direction, the proposed model does not demonstrate promising performance in terms of estimating treatment effects (*cf.*, Section 4). One of the reasons, as acknowledged by the authors, is that CEVAE is not yet capable of addressing the problem of selection bias. Another reason that we think contributes to CEVAE's sub-optimal performance is its graphical model of the underlying data generating mechanism (depicted in Figure 3). This model assumes that there is only one underlying factor Z that generates the entire observational data; however, we know from (Kuang et al., 2017) and (Hassanpour & Greiner, 2020) that there must be more (see Figure 2).

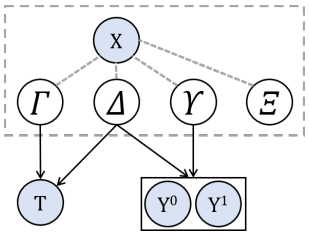

Figure 2. Underlying factors of any observational dataset

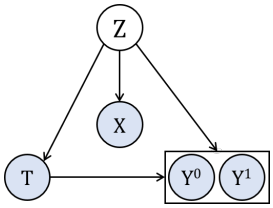

Figure 3. Graphical model of CEVAE (Louizos et al., 2017)

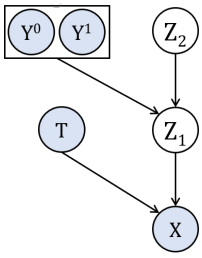

Figure 4. Graphical model of the proposed method

3. Method

Following (Hassanpour & Greiner, 2020) and without loss of generality, we assume that the random variable X follows a(n unknown) joint probability distribution $\Pr(X \,|\, \Gamma, \Delta, \Upsilon, \Xi)$, where Γ, Δ, Υ, and Ξ are non-overlapping factors[3]. Moreover, treatment T follows $\Pr(T \,|\, \Gamma, \Delta)$ and outcome Y^T follows $\Pr_T(Y^T \,|\, \Delta, \Upsilon)$ – see Figure 2. Observe that the factor Γ (resp., Υ) partially determines only T (resp., Y), but not the other variables; and Δ includes the confounding factors between T and Y. This graphical model suggests that selection bias is induced

[1] That is, $X \perp\!\!\!\perp T \,|\, b(X)$ (Rosenbaum & Rubin, 1983)

[2] The authors acknowledge that this would hurt the predictive performance for individual outcomes. As a result, this yields inaccurate estimation of Individual Treatment Effects (ITEs).

[3] As examples: Γ = rich patients receiving the expensive treatment while poor patients receiving the cheap one – although neither of the outcomes depend on patients' wealth status; Δ = younger patients receiving surgery while older patients receiving medication; Υ = genetic information that determines the efficacy of a medication, where this relationship is not known to the attending physician; and Ξ = irrelevant factors such as eye-color.

by factors Γ and Δ. It also shows that the outcome depends on the factors Δ and Υ.

Our goal is to design an architecture for a generative model that encourages disentanglement of these four underlying latent variables (see Figure 2). It is an attempt to decompose and separately learn the underlying factors that are responsible for determining T and Y. Our proposed model employs a VAE (Kingma & Welling, 2014; Rezende et al., 2014) that includes a decoder (generative model) $p_\theta(x|z_1, t)$ and an encoder (variational posterior) $q_\phi(z_1, z_2|x, t, y)$. All components are parametrized as deep neural networks.

Specifically, we use two stacked M2 models (Kingma et al., 2014). The architecture of the proposed method is illustrated in Figure 4. Unlike the M1 model, the M2 model allows the treatment as well as the outcome information to guide the representation learning process. In other words, the proposed structure functions as a distillation tower: the bottom M2 model attempts to decompose Γ (by T) from Δ, Υ, and Ξ (by Z_1); and the top M2 model attempts to decompose Δ and Υ (by Y) from Ξ (by Z_2).

Decoder (parametrized by θ) includes these distributions:

$$\begin{aligned} \text{Priors:} \quad & p_\theta(z_2) \\ & p_\theta(z_1|y, z_2) \\ \text{Likelihood:} \quad & p_\theta(x|z_1, t) \end{aligned}$$

Encoder (parametrized by ϕ) includes these distributions:

$$\begin{aligned} \text{Posteriors:} \quad & q_\phi(z_1|x, t) \\ & q_\phi(y|z_1) \\ & q_\phi(z_2|y, z_1) \end{aligned}$$

The goal is to maximize the conditional log-likelihood of the observed data (left-hand-side of the following inequality) by maximizing the Evidence Lower BOund (ELBO; right-hand-side of the following inequality) – *i.e.*,

$$\begin{aligned} \sum_{i=1}^{N} \log p(x_i|t_i, y_i) \;\geq\; & \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_1|x,t)}\big[\log p_\theta(x_i|z_{1i}, t_i)\big] \\ & - \mathrm{KL}\big(q_\phi(z_1|x, t) \,||\, p_\theta(z_1|y, z_2)\big) \\ & - \mathrm{KL}\big(q_\phi(z_2|y, z_1) \,||\, p_\theta(z_2)\big) \end{aligned}$$

where KL denotes the Kullback-Leibler divergence, $p_\theta(z_2)$ is the unit Gaussian, and the other distributions are parameterized as deep neural networks.

As mentioned earlier, we want the learned latent variables to be disentangled, such that they match to our assumed non-overlapping factors Γ, Δ, Υ, and Ξ. To ensure this, we employ the β-VAE (Higgins et al., 2017) which adds a hyperparameter $\beta > 1$ to the KL part of the ELBO. This adjustable hyperparameter helps balance the latent channel capacity and independence constraints noted by the KL terms) with the reconstruction accuracy, which in turn would encourage disentanglement (Burgess et al., 2018). Therefore, the generative objective to be minimized becomes:

$$\begin{aligned} \mathcal{L}_{\text{VAE}} \;=\; & -\sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_1|x,t)}\big[\log p_\theta(x_i|z_{1i}, t_i)\big] \\ & +\beta \cdot \Big[\mathrm{KL}\big(q_\phi(z_1|x, t) \,||\, p_\theta(z_1|y, z_2)\big) \\ & \quad + \mathrm{KL}\big(q_\phi(z_2|y, z_1) \,||\, p_\theta(z_2)\big)\Big] \end{aligned} \tag{2}$$

Although the proposed graphical model suggests that T and Z_1 are statistically independent (see the collider structure $T \rightarrow X \leftarrow Z_1$ in Figure 4), an information leak is quite possible due to the correlation between the outcome y and treatment t. We therefore require an extra regularization term on the marginal $q_\phi(z_1|t)$ in order to penalize the discrepancy (denoted by `disc`) between conditional distributions of z_1 given $t = 0$ versus given $t = 1$ as follows:

$$\mathcal{L}_{\text{MMD}} \;=\; \texttt{disc}\big(\{z_1\}_{i:t_i=0}, \{z_1\}_{i:t_i=1}\big) \tag{3}$$

Following the literature – *cf.*, (Louizos et al., 2015; Shalit et al., 2017; Hassanpour & Greiner, 2019; 2020) – we use the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) measure to achieve this regularization.

Note, however, that neither the VAE nor the MMD losses contribute to training a predictive model for outcomes. To remedy this, we extend the objective function to include a discriminative term for the regression loss of predicting y. [4]

$$\mathcal{L}_{\text{pred}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \omega_i \cdot \mathcal{L}\big[\, y_i,\, \hat{y}_i\big] \tag{4}$$

where the predicted outcome $\hat{y}_i$ is derived as the mean of the $q_\phi(y_i|z_{1i})$ posterior trained for the respective treatment t_i; $\mathcal{L}\big[\, y_i,\, \hat{y}_i\big]$ is the factual loss (*i.e.*, L2 loss for real-valued outcomes and log loss for binary-valued outcomes); and ω_i represent the weights as derived by Equation (1).

Putting everything together, the overall objective function to be minimized is then:

$$\mathcal{J} \;=\; \mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{MMD}} + \gamma \cdot \mathcal{L}_{\text{VAE}} + \lambda \cdot \mathfrak{Reg}(\,\cdot\,) \tag{5}$$

where $\mathfrak{Reg}(\,\cdot\,)$ regularizes the model complexity. Note that for $\gamma = 0$, the proposed method effectively reduces to CFR-Net. However, as supported by our empirical results (*cf.*, Section 4), the generative term in the objective function helps learning representations that embed more relevant information for estimating outcomes than that of Φ in CFR-Net. We refer to our proposed method as **VAE-CI** (Variational Auto-Encoder for Causal Inference).

[4] This is similar to what is done in (Kingma et al., 2014) by adding a classification loss in their Equation (9).

4. Experiments

4.1. Benchmarks

Infant Health and Development Program (IHDP) The original IHDP randomized controlled trial was designed to evaluate the effect of specialist home visits on future cognitive test scores of premature infants. Hill (2011) induced selection bias by removing a non-random subset of the treated population. The dataset contains 747 instances (608 control and 139 treated) with 25 covariates. We use the same benchmark (100 realizations of outcomes) provided by Johansson et al. (2016) and Shalit et al. (2017).

Atlantic Causal Inference Conference 2018 (ACIC'18) ACIC'18 is a collection of 24 binary-treatment datasets released for a data challenge; with number of instances $N \in \{1, 2.5, 5, 10, 25, 50\} \times 10^3$ (four datasets in each category). The covariates matrix for each dataset is comprised of 177 features and is sub-sampled from a table of medical measurements taken from the Linked Birth and Infant Death Data (LBIDD) (MacDorman & Atkinson, 1998), that contains information corresponding to 100,000 subjects.

4.2. Evaluating Treatment Effect Estimation

Given a synthetic data (that includes both factual and counterfactual outcomes), one can evaluate treatment effect estimation methods with two types of performance measures:

- Individual-based: "Precision in Estimation of Heterogeneous Effect" $\text{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\text{e}}_i - \text{e}_i\right)^2}$ where $\hat{\text{e}}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the predicted effect and $\text{e}_i = y_i^1 - y_i^0$ is the true effect.
- Population-based: "Bias of the Average Treatment Effect" $\epsilon_{\text{ATE}} = \left|\text{ATE} - \widehat{\text{ATE}}\right|$ where $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - \frac{1}{N}\sum_{j=1}^{N} y_j^0$ in which y_i^1 and y_j^0 are the true outcomes and $\widehat{\text{ATE}}$ is calculated based on the estimated outcomes.

4.3. Results and Discussion

In this paper, we compare performances of the following treatment effect estimation methods: **CFR-Net** (Shalit et al., 2017), **Dragon-Net** (Shi et al., 2019), **GANITE** (Yoon et al., 2018), **CEVAE** (Louizos et al., 2017), and **VAE-CI** (our proposed method). We ran the experiments using the publicly available code-bases of the contender methods. Note the following points:

- Since Dragon-Net is designed to estimate ATE only, we did not report its performance results for the PEHE measure (which, as expected, were significantly inaccurate).
- The original GANITE code-base was implemented for binary outcomes only. We modified the code (losses, etc.) such that it could process real-valued outcomes also.

Table 1. PEHE and ϵ_{ATE} on the IHDP (100 realizations) dataset

METHOD	PEHE	ϵ_{ATE}
CFR	0.70 ± 0.36	0.07 ± 0.08
DRAGON	NA	0.14 ± 0.12
GANITE	5.11 ± 7.85	1.12 ± 2.28
CEVAE	2.50 ± 3.48	0.18 ± 0.25
VAE-CI	0.60 ± 0.19	0.02 ± 0.03

Table 2. PEHE and ϵ_{ATE} on the ACIC'18 ($N \leq 10K$) dataset

METHOD	PEHE	ϵ_{ATE}
CFR	4.46 ± 7.97	1.21 ± 1.98
DRAGON	NA	0.95 ± 1.68
GANITE	5.06 ± 5.81	1.30 ± 1.85
CEVAE	5.65 ± 6.71	2.87 ± 3.44
VAE-CI	1.82 ± 2.08	0.79 ± 1.50

- We were surprised that CEVAE diverged when running on the ACIC'18 datasets. To avoid this, we had to run the ACIC'18 experiments on the binary covariates only.

Table 1 summarizes the mean and standard deviation of the PEHE and ϵ_{ATE} measures (lower is better) on the IHDP benchmark. VAE-CI achieves the best performance among the contending methods (statistically significant based on the Welch's unpaired t-test with $\alpha = 0.05$).

Table 2 reports the PEHE and ϵ_{ATE} measures on the 16 datasets in the ACIC'18 benchmark with $\leq 10K$ samples. Similar to the IHDP benchmark, VAE-CI achieves the best performance among the contending methods. These results, however, are not statistically significant, which is mostly due to the high standard deviation of the contending methods.

5. Future works and Conclusion

We hypothesize that the most important reasons for the superior performance of the proposed method are: (i) the architecture of our double-stacked M2 VAE model; and (ii) the disentanglement property of the β-VAE component. More theoretical and empirical analysis should be conducted to support this claim though, which is left to future work.

In this paper, we employed a variant of the semi-supervised VAE (Kingma et al., 2014) in a novel double-stacked M2 architecture in order to estimate treatment effects (for both individuals as well as the entire population). For the VAE component, we used β-VAE (Higgins et al., 2017) to encourage learning disentangled representations. Our empirical results demonstrated the superiority of the proposed method compared to both state-of-the-art discriminative as well as generative approaches in the literature.

References

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 13 (March), 2012.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, 2019.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *ICLR*, 2020.

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

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *ICML*, 2016.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014.

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

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *NeurIPS*. 2017.

MacDorman, M. F. and Atkinson, J. O. Infant mortality statistics from the 1996 period linked birth/infant death dataset. *Monthly Vital Statistics Report*, 46(12), 1998.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 1974.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*, 2019.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, 2018.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.