	Setting A		Setting B	
Methods	$\epsilon^{tr}_{\text{PEHE}}$	$\epsilon^{te}_{\text{PEHE}}$	$\epsilon^{tr}_{\text{PEHE}}$	$\epsilon^{te}_{\text{PEHE}}$
CT	1.48±0.12	1.56±0.13	5.46±0.08	5.73±0.09
t-stats	1.78±0.09	1.91±0.12	5.40±0.08	5.71±0.09
CF	1.01±0.08	1.09±0.16	3.86±0.05	3.91±0.07
BART	0.87±0.07	0.88±0.07	2.78±0.03	2.91±0.04
X-RF	0.98±0.08	1.09±0.15	3.50±0.04	3.59±0.06
CFR	0.67±0.02	0.73±0.04	2.60±0.04	2.76±0.04
SITE	0.65±0.07	0.67±0.06	2.65±0.04	2.87±0.05
DR-CFR	0.62±0.15	0.65±0.18	2.73±0.04	2.93±0.05
GANITE	1.84±0.34	1.90±0.40	3.68±0.38	3.84±0.52
CEVAE	0.95±0.12	1.04±0.14	2.90±0.10	3.24±0.12
TEDVAE	**0.59±0.11**	**0.60±0.14**	**2.10±0.09**	**2.22±0.08**

Table 2: Means and standard deviations of the PEHE metric (smaller is better) on IHDP. The bolded values indicate the best performers (Wilcoxon signed rank tests ($p = 0.05$).

	Twins	
Methods	$\epsilon^{tr}_{\text{ATE}}$	$\epsilon^{te}_{\text{ATE}}$
CT	0.034±0.002	0.038±0.007
t-stats	0.032±0.003	0.033±0.005
CF	0.025±0.001	0.025±0.001
BART	0.050±0.002	0.051±0.002
X-RF	0.075±0.003	0.074±0.004
CFR	0.029±0.002	0.030±0.002
SITE	0.031±0.003	0.033±0.003
DR-CFR	0.032±0.002	0.034±0.003
GANITE	0.016±0.004	0.018±0.005
CEVAE	0.046±0.020	0.047±0.021
TEDVAE	**0.006±0.002**	**0.006±0.002**

Table 3: Means and standard deviations of ϵ_{ATE} on the Twins datasets. The bolded values indicate the best performers (Wilcoxon signed rank tests ($p = 0.05$).

the outcomes follow linear relationship with the variables in "Setting A" and exponential relationship in "Setting B". The datasets can be accessed at https://github.com/vdorie/npci. The reported performances are averaged over 100 replications with a training/validation/test splits proportions of 60%/30%/10%.

Since evaluating treatment effect estimation is difficult in real-world scenarios (**?**), a good treatment effect estimation algorithm should perform well across different datasets with minimum requirement for parameter tuning. Therefore, for TEDVAE we use the same parameters in the ACIC dataset and do not perform parameter tuning on the IHDP dataset. For the compared traditional methods, we also use the same parameters as selected on the ACIC benchmark. For the compared deep learning methods, we conduct grid search using the recommended parameter ranges from the relevant papers.

From Table 2 we can see that TEDVAE achieves the lowest PEHE errors among the compared methods on both Setting A and Setting B of the IHDP benchmark. Wilcoxon signed rank tests ($p = 0.05$) indicate that TEDVAE is significantly better than the compared methods. Since TEDVAE uses the same parameters on the IHDP datasets as in the previous ACIC benchmarks, these results demonstrate that the TEDVAE model is suitable for diverse real-world scenarios and is robust to the choice of parameters.

Real-world Dataset: Twins In this section, we use a real-world randomized dataset to compare the methods capability of estimating the average treatment effects.

The Twins dataset has been previously used for evaluating causal inference in (**??**). It consists of samples from twin births in the U.S. between the year of 1989 and 1991 provided in (**?**). Each subject is described by 40 variables related to the parents, the pregnancy and the birth statistics of the twins. The treatment is considered as $t = 1$ if a sample is the heavier one of the twins, and considered as $t = 0$ if the sample is lighter. The outcome is a binary variable indicating the children's mortality after a one year follow-up period. Following the procedure in (**?**), we remove the subjects that are born with weight heavier than 2,000g and those with missing values, and introduced selection bias by removing a non-random subset of the subjects. The final dataset contains 4,813 samples. The data splitting is the same as previous experiments, and the reported results are averaged over 100 replications. The ATE estimation performances are illustrated in Table **??**. On this dataset, we can see that TEDVAE achieves the best performance with the smallest ϵ_{ATE} among all the compared algorithms.

Overall, the experiments results show that the performances of TEDVAE are significantly better than the compared methods on a wide range of synthetic, benchmark, and real-world datasets. In addition, the results also indicate that TEDVAE is less sensitive to the choice of parameters than the other deep learning based methods, which makes our method attractive for real-world application scenarios.

Conclusion

We propose the TEDVAE algorithm, a state-of-the-art treatment effect estimator which infer and disentangle three disjoints sets of instrumental, confounding and risk factors from the observed variables. Experiments on a wide range of synthetic, benchmark, and real-world datasets have shown that TEDVAE significantly outperforms compared baselines. For future work, a path worth exploring is extending TEDVAE for treatment effects with non-binary treatment variables. While most of the existing methods are restricted to binary treatments, the generative model of TEDVAE makes it a promising candidate for extension to treatment effect estimation with continuous treatments.

Eveniet ea ex illum nam incidunt quas sint quo quod explicabo, debitis dolor odio quibusdam perspiciatis, placeat officia quae animi error, provident optio consectetur voluptatum reiciendis, delectus saepe deleniti excepturi repellat molestias vel ut?At laboriosam iste ullam repudiandae id ea eveniet distinctio aut eos, laudantium corporis veniam, cum culpa esse cupiditate dicta ullam unde rem minima inventore numquam, quis doloremque ea, laudantium sequi eius aliquam esse officia in?Iusto optio ad sed asperiores numquam corrupti delectus totam neque qui, modi molestiae laborum

aliquid earum vel rerum sapiente, voluptates illum necessitatibus optio, cupiditate in voluptatem unde aliquam aliquid at.