The robustness of GANITE to hidden confounders

Vincent C.O. van Oudenhoven

Supervisors: Stephan Bongers, Jesse Krijthe

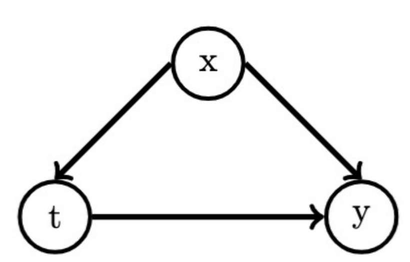
Faculty of Electrical Engineering, Mathematics and Computer Science

Background

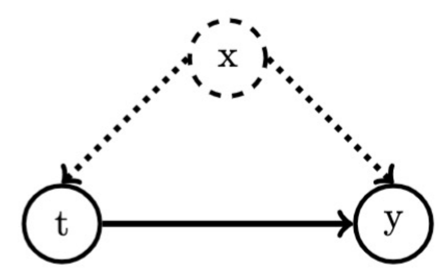
- The causal inference branch of Causal Machine Learning (CML)
  - Treatment effect estimation
- A confounder, **x**, has a confounding effect through the treatment, **t**, and the outcome, **y**
  - The link from **x** → **t** can be represented as a "propensity score"
- Rubin-Neyman model as a foundation with key assumptions [1]
  - Unconfoundedness: all is measured
  - Overlap: non-zero propensity of all treatments
- Treatment effects may be homogenous or heterogenous
- The treatment effects can be estimated for different subgroups

$$\mathrm{ITE}_i = y_i^1 - y_i^0 \quad (1)$$

$$\mathrm{ATE} = \frac{1}{n}\sum_{n=1}^{n} \mathrm{ITE}_i \quad (2)$$



Confounder, x, on t and y



Hidden confounder, x, on t and y

Research question

How robust is GANITE to hidden confounders?

- What happens to the performance of the model and the inferred ATE when single confounders are removed?
- How does GANITE behave as more confounders are removed?

Methodology

- The model under test is called "Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE)" [2]
  - To model the distribution of counterfactuals indirectly
  - Leverages two distinct Generative Adversarial Networks (GANs)
  - Optimized through the estimated counterfactuals
- The model is tested on three datasets:
  - Infant Health and Development Program (IHDP) [3]
  - Twins [4]
  - Synthetic data
- The performance is evaluated through:
  - Precision in Estimating Heterogenous Effects (PEHE) [5]
  - Deviation in inferred ATE from ground truth

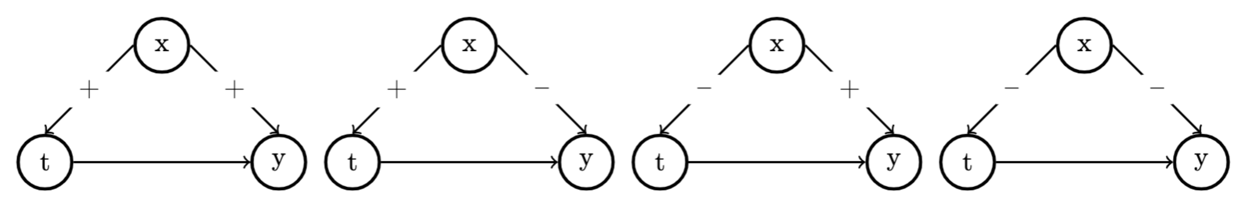
$$\sqrt{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i^1 - y_i^0\right) - \left(\widehat{y_i^1} - \widehat{y_i^0}\right)\right)^2} \quad (3)$$

- The synthetic dataset is generated with known causal graphs
  - Fixed causal strength on treatment and outcome
  - Heterogenous treatment effect

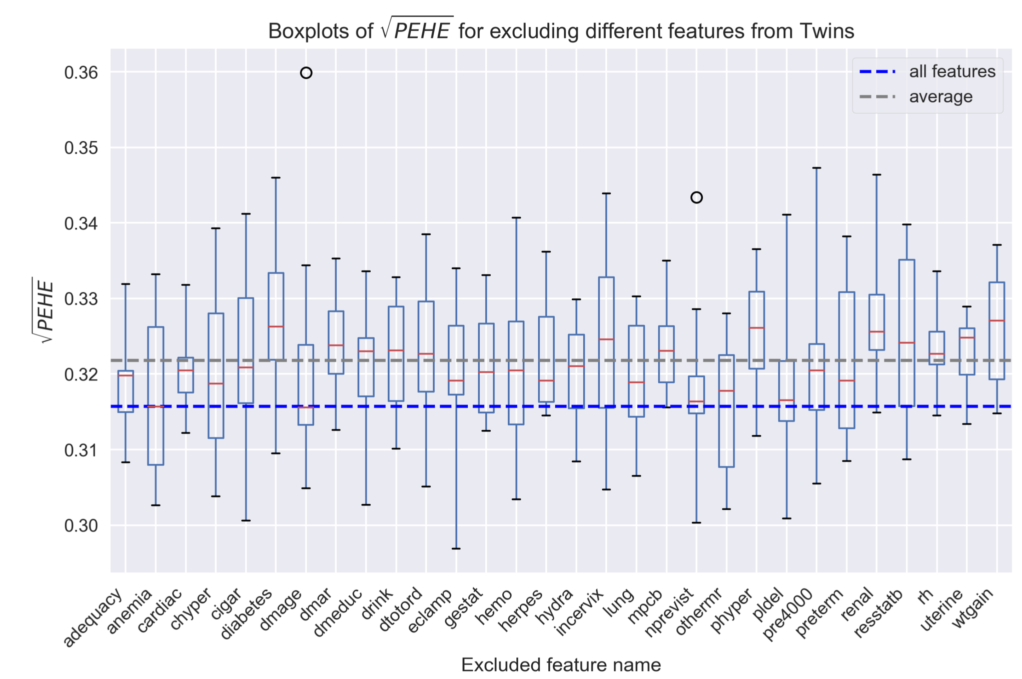
Single feature removal

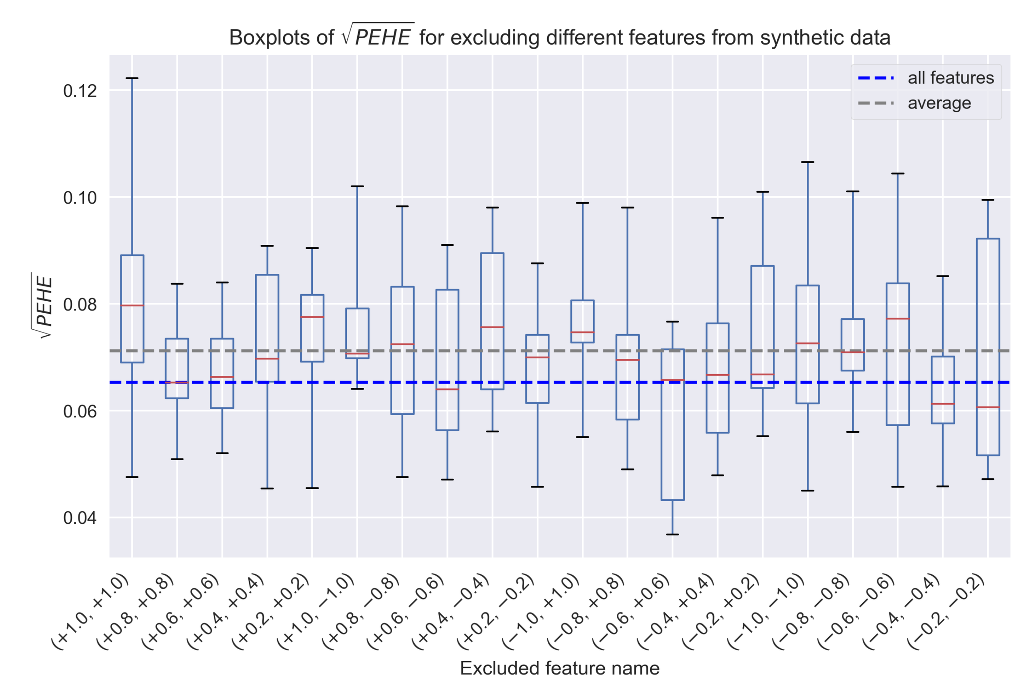
Hypotheses:

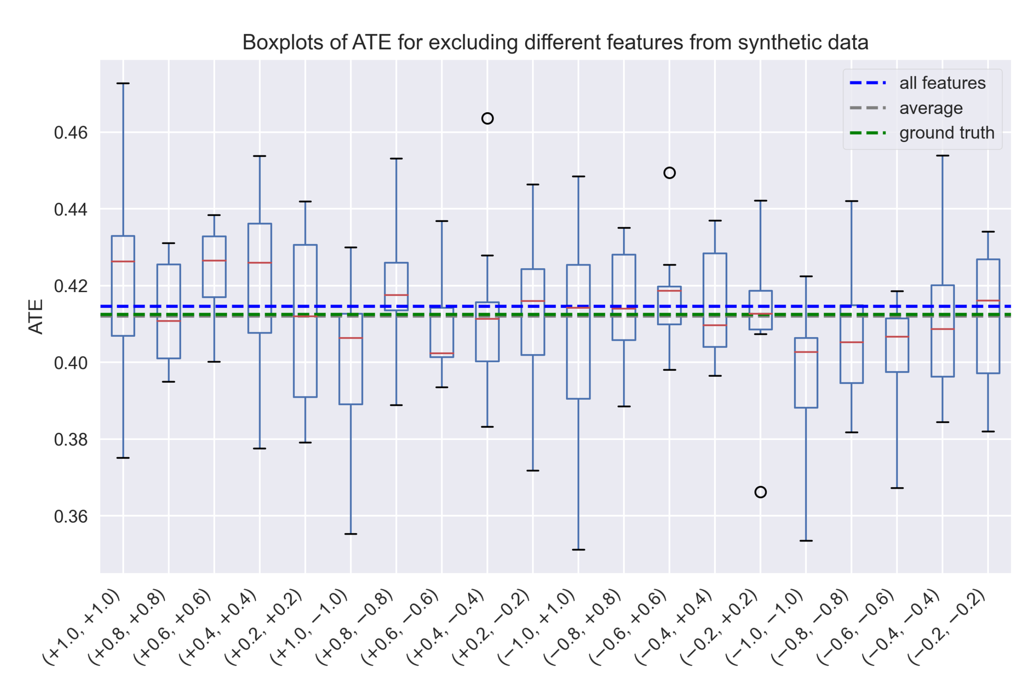
1. The $\sqrt{PEHE}$ will increase relative to the feature's causal effects. On average the error should be higher.
2. The inferred ATE can increase or decrease based on the removed confounder's causal graph.



Kinds of confounders, categorized by their polarity



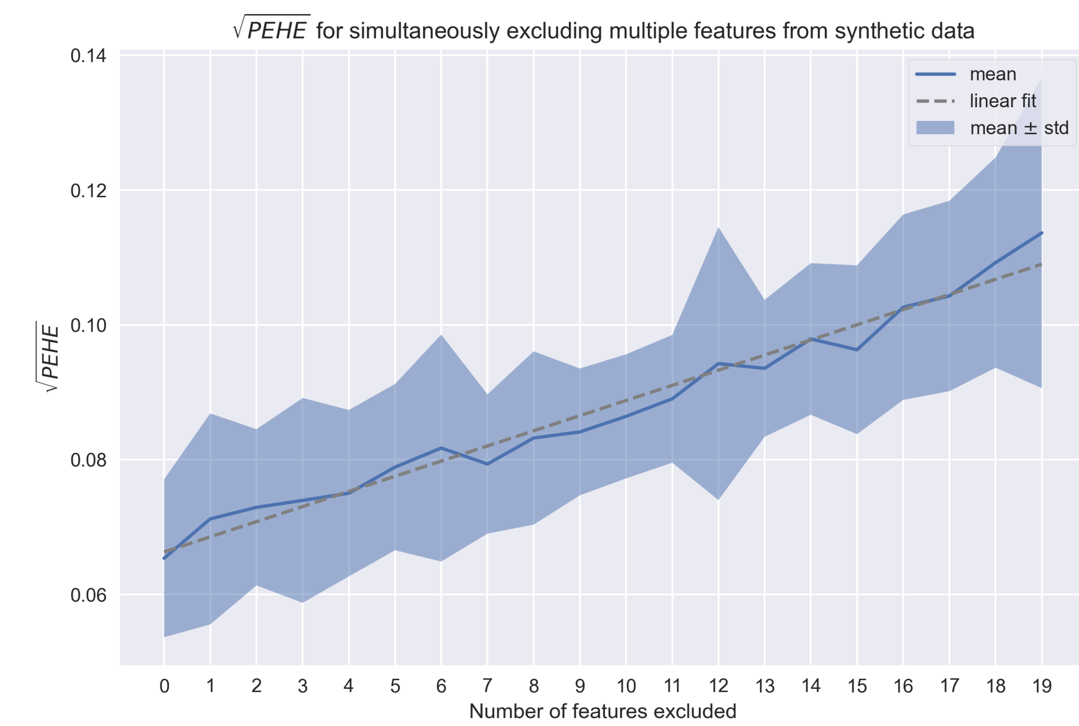


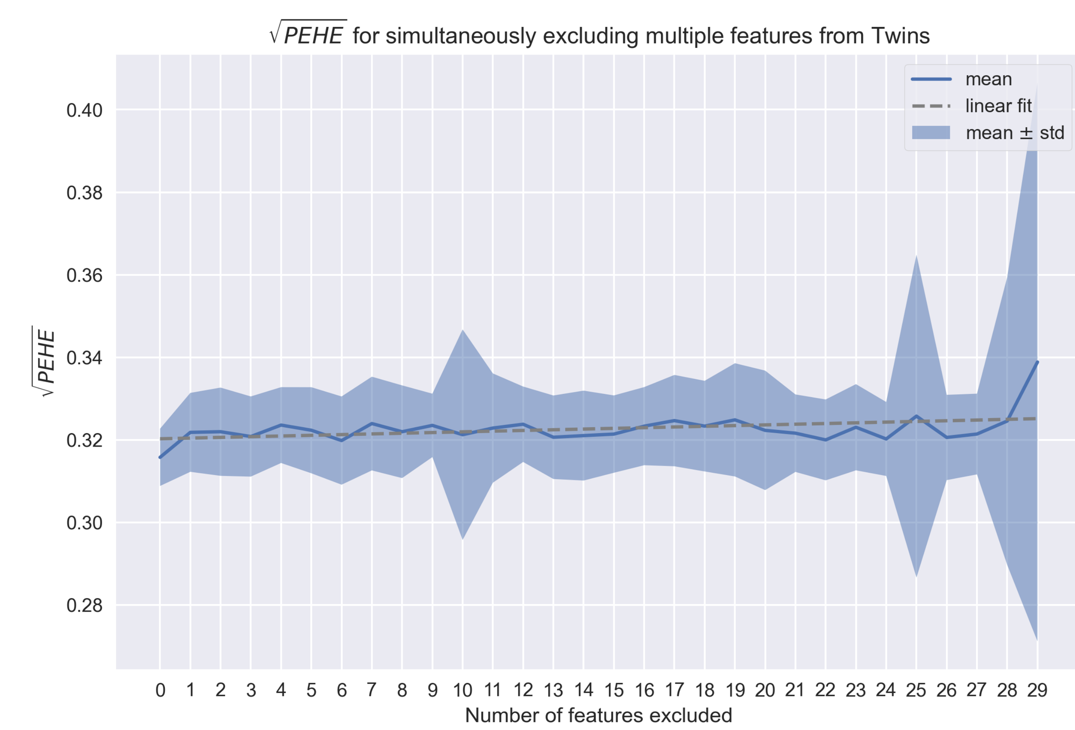


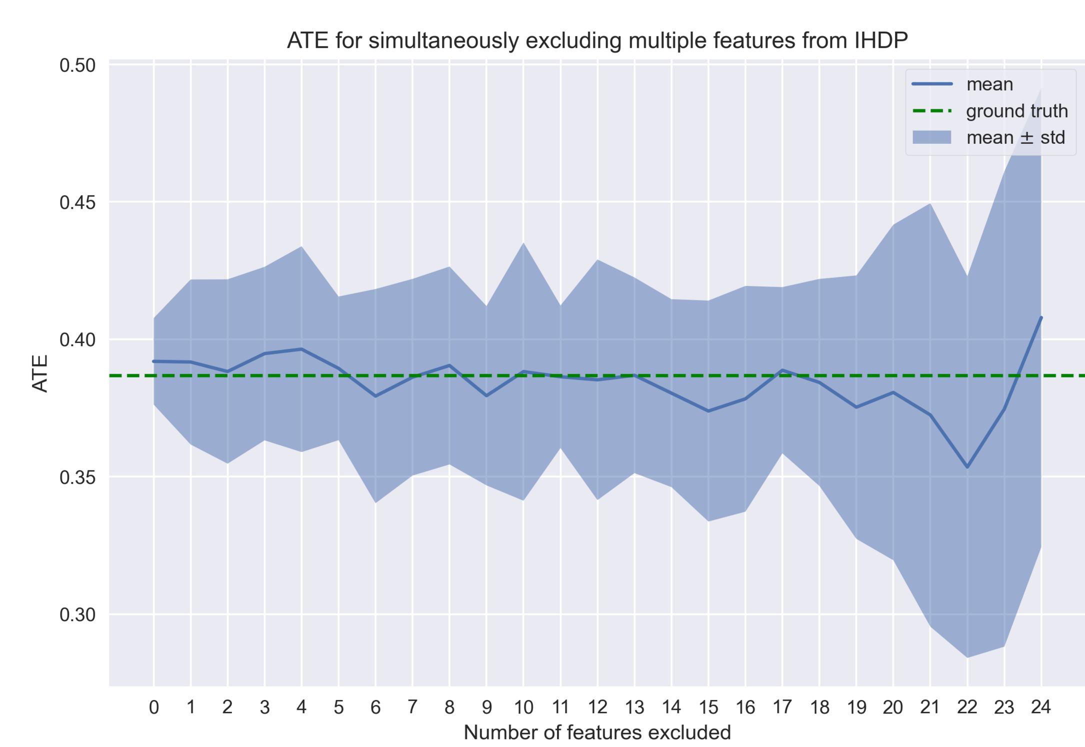
Simultaneous feature removal

Hypotheses:

1. As more features are removed the $\sqrt{PEHE}$ will increase
2. Assuming datasets have are balanced in terms of types of confounders there should be no change to the inferred ATE
3. The variability of the predictions will increase as more confounders are hidden







Conclusion

- Upon hiding single confounders, the $\sqrt{PEHE}$ increases and the inferred ATE can vary from the ground truth
- The effect of the causal strength of the removed confounders on the error is not apparent
- There is no clear pattern for the inferred ATE relative to the ground truth for different kinds of confounders
- For IHDP and the synthetic dataset the $\sqrt{PEHE}$ and the number of hidden confounders are positively correlated
  - The error metrics for Twins contrast this
- The variability of predictions goes up as more confounders are removed
- Tuning hyper-parameters is costly and can lead to biased results
- GANITE is generally hard to train, and its instability is a limiting factor to this study.
  - Likely due to the underlying GANs [6]

Future Work

- Experiments should be repeated with many more trials
- Exploration of the trend in variance of predictions under confounders
- A metric that quantifies the importance of a feature based on its causal graph.
- Comparison to other ITE methods
- More complex synthetic data could be used to better simulate real world data
  - Underlying distributions such as uniform and exponential
  - Nonlinear functions for causal effect
- Optimize the hyper-parameters for each experiment
- Exploring the overlap assumption

References

[1] Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association

[2] Yoon, J., Jordon, J., and van der Schaar, M. (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. In International Conference on Learning Representations.

[3] Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms.

[4] Almond, D., Chay, K. Y., and Lee, D. S. (2005). The Costs of Low Birth Weight. The Quarterly Journal of Economics.

[5] Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics.

[6] Saxena, D. and Cao, J. (2022). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. ACM Computing Surveys.

Links

Email: vincentvanoudenhoven@gmail.com

Repository: https://github.com/vcovo/cse3000