Empirical study of GANITE's robustness to hidden confounders

Vincent C.O. van Oudenhoven
Supervisor(s): Stephan Bongers, Jesse Krijthe
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

An empirical study is performed exploring the sensitivity to hidden confounders of GANITE, a method for Individualized Treatment Effect (ITE) estimation. Most real world datasets do not measure all confounders and thus it is important to know how crucial this is in order to obtain comparable predictions. This is explored through the removal of confounders with varying strengths and by removing subsets of the confounders simultaneously. The sensitivity is measured through the change in Precision in Estimating Heterogeneous Effects (PEHE) and through the divergence in the estimation of Average Treatment Effect (ATE) from the GT. Experiments are performed on synthetic and semi-synthetic data. The number of removed hidden confounders increases the error and variability of predictions, both for ITE and ATE. The strength of the removed confounders does not show a conclusive relationship on the error metrics. The effect of removing confounders with different causal graphs is explored but fails to show any clear patterns due to the high variance of the results.

1 Introduction

A particular problem within artificial intelligence is that of causality: namely does one thing cause another or are they only correlated? The field exploring this is known as Causal Machine Learning (CML). In recent work Richens et al. (2020) has shown promise that such methods can drastically improve the accuracy of medical diagnoses. The two main problems that CML concerns itself with, as per van der Schaar and Maxfield (2021), are: (i) causal discovery: namely finding what variables affect others and in what direction and (ii) causal effect estimation: finding the strength of these effects between variables, the focus tends to be on determining the strength of the effect of a treatment on the outcome. This study focuses on the second problem. Causal prediction is different from traditional machine learning inference as it doesn't only try to predict an outcome but what the effect of a treatment is on that outcome. For example the question may change from "will a patient live?" to "will a patient live because of a particular prescription?". In order to be able to determine the effect of a treatment on an individual, models generally need to learn from data. Data that is frequently incomplete - it is important to know how much this matters to the reliability of the predictions.

In order to understand the rest of this paper it is important to define some of the main terminology. When studying the causal effect of one variable, $\mathbf{t}$, on another, $\mathbf{y}$, a "confounder" is a third variable, $\mathbf{x}$, that has a causal effect on both $\mathbf{t}$ and $\mathbf{y}$, as depicted in Figure 1. This means that when estimating the effect of $\mathbf{t}$ on $\mathbf{y}$ one must condition on $\mathbf{x}$. This effect is called the "treatment effect" where for this example $\mathbf{t}$ would be the treatment. This treatment effect can be estimated for different subsets of the population, with the extremes being: for single individuals or for the entire population. The former is known as Individualized Treatment Effect (ITE), shown in equation Equation 1 for a sample i, and the latter is known as Average Treatment Effect (ATE), shown in Equation 2 Hill (2011). Other measurements exist for subgroups of the population but are not of importance for this study. Both the ITE and ATE are important to infer correctly and that's why in this study the performance of the model is quantified both through an average error over ITE, which is made concrete in subsection 2.4, and the inferred ATE and how this compares to the ground truth, denoted GT.

$$\text{ITE}_i = y_i^1 - y_i^0 \tag{1}$$

$$\text{ATE} = \frac{1}{n}\sum_{n=1}^{n} \text{ITE}_i \tag{2}$$

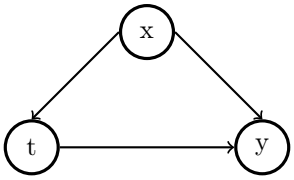


Figure 1: Example of a confounder, $\mathbf{x}$, on $\mathbf{t}$ and $\mathbf{y}$

There are numerous approaches to tackle causal effect estimation, as described by Guo et al. (2021). The specific method explored in this research is called Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) developed by Yoon et al. (2018). Estimating ITEs is part of causal effect estimation and essentially tries to determine the treatment effect based on an individual's features. In order to infer ITEs: GANITE uses Generative Adversarial Nets (GANs). A GAN is composed of a generator and a discriminator where "the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution" (Goodfellow et al., 2014, p. 1). These two models compete until the distribution of the generated samples converges to the distribution of the real data. This is exactly the mechanism that motivated the creation of GANITE, namely to be able to estimate the distribution of the counterfactuals which are not known for the majority of real world datasets, in order to better estimate the ITEs. Details of the architecture of GANITE are provided in subsection 2.1.

GANITE, similarly to many other models that are built on the Rubin-Neyman model, first brought forward by Rubin (2005), rely on this model's underlying assumptions. The particular assumption of interest is that of "unconfoundedness" meaning that all confounding variables are accounted for. More concretely this means that when conditioning on the feature vector, $\mathbf{x}$, the potential outcomes, $\mathbf{y}$, and the treatment, $\mathbf{t}$ are independent; the joint influences have been accounted for through $\mathbf{x}$. Unfortunately this assumption rarely holds with real world datasets where it is practically impossible to measure all influences on the outcome and the treatment. Thus in order to use GANITE in practice, on data with hidden confounders, it is important to know what the effects are on performance when violating this assumption.

In order to shed light on the usability of GANITE this work aims to answer the research question: "How robust is GANITE to hidden confounders?". More specifically what happens to the performance of the model and the inferred ATE when single confounders are removed? Furthermore, how does GANITE behave as more confounders are removed - how much of the dataset can be removed before the predictions become unreliable? By answering these questions the usefulness of GANITE can be further established for the application on real world datasets where hidden confounders are practically inevitable. The results may also provide insight into conditions when GANITE stops performing well altogether and is better avoided. The hypotheses together with the experiments that aim to test them are presented in subsection 2.2. Experiments are performed on three datasets, two of which are semi-synthetic and one is entirely generated for this study. The results are brought into perspective by comparing to the error obtained when training on AF and by comparing the inferred ATE with the known GT for each dataset. All in all, the error and variance in the predictions increases as confoundedness is increased while no conclusive effects are found on the inferred ATE under different kinds of hidden confounders.

The problem setting and how the quantitative analysis is performed with the respective experimental setup are presented in section 2. The results of each experiment together with an analysis in the light of the hypotheses is given in section 3 and section 4. The findings are used to answer and conclude upon the research questions in section 5. In the final section of the paper, section 6, open issues, limitations, possible improvements and extensions, are discussed. After the main paper,

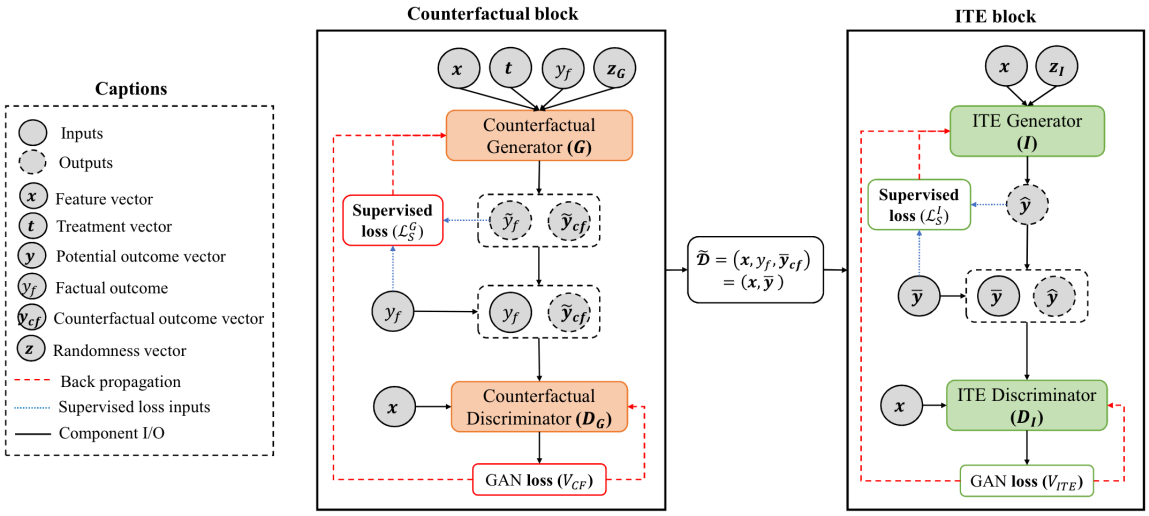


Figure 2: GANITE architecture overview by Yoon et al. (2018)

section 7, is dedicated to exploring the ethical implications of the research, the reproducibility of the results and the integrity of the work as a whole.

2 Methodology

This section describes the GANITE model together with the decisions made when testing it followed by overviews of the datasets used for this study where the generation of the synthetic dataset is presented in detail. This is followed by how the model is evaluated and what experiments are set up to do so.

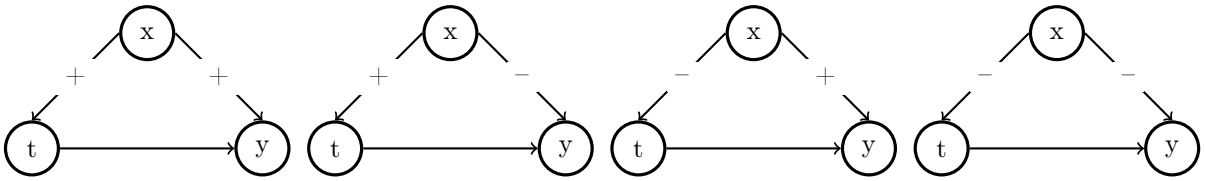
2.1 GANITE

GANITE is composed of two separate GANs as seen in Figure 2. The first GAN, on the left of the figure, concerns itself with the generation of counterfactuals, namely the potential outcomes that are not observed. This is done by training on $\mathbf{x}$, $\mathbf{t}$, and the factual outcomes, $\mathbf{y_f}$ in order to learn the distribution of the counterfactuals. This is learnt through back propagating the error from the counterfactual discriminator which, in a nutshell, assigns a error relative to the perceived likelihood that that sample came from the original distribution. A "complete dataset" can then be constructed by joining the original data and the these generated counterfactuals. The second GAN, on the right of the figure, is trained on this complete dataset and is used for the inference of the ITEs. This second GAN's generator is used during inference while all components of both GANs are used during training.

There are numerous hyper-parameters that can be adjusted for GANITE. The hyper-parameters are dataset dependent but considering that finding the optimum for each experiment implies increasing the search space by numerous orders of magnitude this was deemed infeasible for this study and thus for each dataset they are kept constant throughout the trials. The values used for the experiments are based on the proclaimed optimal hyper-parameters, described in (Yoon et al., 2018, p. 15), for the IHDP and Twins datasets. As described in subsection 2.3 the synthetic dataset is most similar to IHDP in terms of it's meta characteristics and thus the same hyper-parameters as IHDP were used for these experiments. The batch size controls how many samples are fed at a time to

Hyper-parameter	Twins value	IHDP value	Synthetic value
Batch size	128	64	64
Hidden dimensions	8	8	8
Alpha	2	2	2
Iterations	2000	1000	1000

Table 1: Hyper-parameters used for GANITE per dataset



(a) Positive on **t** and positive on **y**

(b) Positive on **t** and negative on **y**

(c) Negative on **t** and positive on **y**

(d) Negative on **t** and negative on **y**

Figure 3: Possible polarity combinations of the effect of confounder **x** on **t** and **y**

the GANs, the number of hidden dimensions determines the dimensionality of the hidden layer(s), alpha controls the importance of the loss and finally the iterations is how many times the dataset is fed into the GANs until the training is stopped. Refer to the original paper for further details.

2.2 Experimental Setup

In order to answer the research question the experiments have been split up into two parts: one where single features are removed and one where multiple are removed simultaneously. Each experiment is repeated multiple times in order to account for the variability and randomness inherent to GANs. For each trial the dataset is split up randomly using a 80/20 train and test split.

A confounder's causal effects can take on various forms aside from the magnitude itself. Restricting this study to the simple case of confounders depicted in Figure 1 the combinations depicted in Figure 3 can be obtained. It is assumed that the treatment effect on the outcome is positive. For example the (+, +) is a confounder, shown in Figure 3a, that increases the likelihood of treatment while having a positive effect on the outcome: this can be the characteristic of being "health conscious" and thus being more likely to reach out for a treatment while also being more likely to have a good outcome due to numerous other healthy choices that come from this characteristic. Upon the removal of this confounder it is no longer possible to account for it's causal effects and the treatment will seem to have a stronger treatment effect for those individuals. The "importance" of a confounder is defined as the magnitude of the causal graphs on the treatment and outcome.

2.2.1 Single feature removal

When hiding single confounders at a time the hypotheses are:

1. The $\sqrt{PEHE}$ will increase relative to the feature's importance as removing it will obfuscate more information to the model.

2. The average $\sqrt{PEHE}$ over all trials with a removed feature should be higher than the $\sqrt{PEHE}$ of AF because without the ability to account for any single confounder the model should have less success with the inference of the ITE.

3. The ATE can increase or decrease based on the feature's causal graph on the treatment and the outcome, for which the following hypotheses are posed:

   (a) Upon the removal of a (+, +) confounder the inferred ATE will go up because the positive outcome will be credited to the treatment instead of the hidden confounder, which is further enhanced by the increased likelihood of treatment.

   (b) Upon the removal of a (+, −) confounder the inferred ATE will go down: for example individuals with a mortal disease may be more likely to be treated while also being more likely to die after the treatment.

   (c) Upon the removal of a (−, +) confounder the inferred ATE will go down: as a healthy individual may be less likely to be treated while having a higher chance of survival after the treatment, thus the ATE will be underestimated.

   (d) Upon the removal of a (−, −) confounder the inferred ATE will go up because of the same but inverted reason as for (+, +).

In order to test these hypotheses single features are removed one at a time for all features, repeated 10 times each. The $\sqrt{PEHE}$ and inferred ATE are measured and compared to the error obtained when training over all features and to the GT ATE, respectively.

2.2.2 Simultaneous feature removal

The effect of multiple hidden confounders on GANITE is tested through simultaneous removal of features from the dataset. In order to quantify the effect of removing a certain number of features it is crucial to take an average over multiple possible combinations. The number of combinations of removing n features from k is quantified by $\binom{n}{k}$ which grows proportionately to $O(n^k)$ and thus testing all combinations is computationally unfeasible. Instead the minimum of 50 trials and the number of possible combinations (e.g. there are only 20 ways of removing 19 features from 20) is used as an approximation. The variability of the trials for each experiment is quantified through standard deviation.

The hypotheses for these experiments are:

4. The $\sqrt{PEHE}$ will increase as more features are removed as the model is less able to differentiate individuals.

5. There is little to no change in the inferred ATE as more features are removed as the datasets are assumed to be comprised of balanced confounders.

6. The variability of the estimations will increase due to the model having less information to learn from and thus being less likely to converge to the same state.

2.3 Datasets

In order to evaluate the performance of GANITE two types of data have been used. The starting point for this study is semi-synthetic data where the factuals and counterfactuals are known, described in subsubsection 2.3.1 and subsubsection 2.3.2. In these datasets the ITE can be calculated directly from the potential outcomes and the GT ATE can be derived from the ITEs. The only drawback of these datasets is that the causal graphs relating the features to the treatment and the outcome are unknown. For this reason a second kind of data was created specifically for this study, described in subsubsection 2.3.3. In order to limit the size of the study only binary treatments are considered. During the experimentation phase it became clear that GANITE was clipping the outputs to be between 0 and 1. This was no problem for Twins as the outcomes are already binary

but it was required to min max scale, shown in Equation 3, both IHDP and the synthetic dataset. This avoids any functional changes to GANITE that may have adverse effects.

$$\text{min\_max\_scaling}(y) = \frac{y - y_{min}}{y_{max} - y_{min}} \tag{3}$$

2.3.1 IHDP

The Infant Health and Development Program (IHDP) dataset is described by (Hill, 2011, p. 10) as originating from a randomized experiment where premature infants were either treated with intense child care or not. The features include measurements of the child, the physical and socioeconomic conditions of the mother and the residence of the family. The specific replication used for this study comes from later work done by Shalit et al. (2017) where the dataset has "been made imbalanced by removing a biased subset of the treated population" - this is appealing for this study as it increases the complexity of the dataset and is thus more representative of real world datasets that frequently do not come from randomized controlled trials. The motivation behind using an IHDP replication is because Guo et al. (2021)'s survey considers it a benchmark dataset for causal inference models. Unfortunately due to the number of times that this dataset has been passed along, for differing purposes, the mapping of the features of the used dataset and those described by Hill (2011) have been obfuscated. It is because of this that no additional insights into why a higher error or deviation in ATE is observed for particular features can be provided.

2.3.2 Twins

The Twins dataset is a semi-synthetic dataset consisting of all twin births in the USA between 1989-1991, totalling 11,400 pairs of twins, described by Almond et al. (2005). The dataset has a total of 30 features that encompass characteristics of the parents, the pregnancy and birth, the received care, and the residence of the twins. For this dataset the treated is defined as the heavier twin. This is a popular dataset for CML because there are always two twins and all of their features aside from the treatment are equal and thus the factuals and counterfactuals are known. The causal relationships of the features on the treatment and outcome are not known although their strength can be estimated to a certain degree based on medical evidence.

2.3.3 Synthetic data

In order to test the hypotheses the individual effects need to be known in addition to the causal effects between the features and the treatment and output. The latter is not known for both IHDP and Twins and thus a synthetic dataset was created. The synthetic dataset used for this study consists of a total of 20 confounders, a binary treatment and a continuous outcome. The confounders are chosen such that each category is equally represented: both in the number of confounders and in the causal strength of those confounders on the treatment and outcome. This is done for two reasons: the first is that upon removing e.g. a $(+, +)$ confounder the observed phenomena are within the context of having other kinds of confounders and is thus not dependent on different dataset conditions for each confounder category. The second reason is that this is more representative of a real world dataset where it is likely that there are confounders of each kind. The magnitude of the causal effect of the confounder is kept the same on the treatment and the outcome. Knowing the causal graphs allows for the comparison of the effect of hiding a particular confounder to its category and strength. For this dataset the causal effect of the confounder on the treatment is modelled as a propensity score, namely the likelihood that an individual will be treated, as introduced by Rosenbaum and Rubin (1983).

The treatment effect, treatment propensity, and outcome all have a degree of noise, small relative to the values themselves. The strengths of the confounders are predefined. The values for the confounders are sampled from a normal distribution. The propensity score is a min max scaled weighted sum of the features - in other words high feature values lead to a higher chance of treatment. Finally, due to GANITE's outcome restrictions, $\mathbf{y}$ is min max scaled. The dimensionality is represented by j and the sample number is represented by i. The total number of samples is denoted n.

$$ITE_{(\text{noise},i)}, \text{propensity}_{(\text{noise},i)}, y_{(\text{noise},i)} \sim N(0, 0.01) \tag{4}$$

$$x_{(\text{causal strengths},i,j)} \in (\pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1.0) \tag{5}$$

$$x_{(\text{values},i,j)} \sim N(1, 1) \tag{6}$$

$$\text{propensity}_i = \text{min\_max\_scaling}(\sum_{j=1}^{n} x_{(\text{causal strengths},i,j)} * x_{(\text{value},i,j)} + \text{propensity}_{(\text{noise},i)}) \tag{7}$$

$$t_i \sim Ber(\text{propensity}_i) \tag{8}$$

$$ITE_i = \sum_{j=1}^{n} x_{(\text{value},i,j)} + ITE_{(\text{noise},i)} \tag{9}$$

$$y_i^0 = \sum_{j=1}^{n} x_{(\text{causal strengths},i,j)} * x_{(\text{value},i,j)} + y_{\text{noise}} \tag{10}$$

$$y_i^1 = y_i^0 + ITE_i \tag{11}$$

$$y_i = \text{min\_max\_scaling}([y_i^0, y_i^1]) \tag{12}$$

2.4 Model evaluation

In order to quantify the effect of each experiment two metrics are of concern. First off, as GANITE is designed for ITE the most important metric for this work is the Precision in Estimation of Heterogeneous Effect (PEHE), shown in equation 13, which quantifies the squared difference between the estimated outcome and the GT for each individual. The square root of this metric is taken in order for the units to be directly comparable with the estimations of treatment effect, which is also done by Yoon et al. (2018) during the original evaluation of GANITE - this is analogous to the root mean squared error of the ITE. $\sqrt{PEHE}$ is commonly used for the evaluation of ITE error such as by Hill (2011), Shalit et al. (2017), and Louizos et al. (2017). The second metric used for evaluation is the inferred ATE relative to both the GT and the ATE inferred, as described by Equation 2. All metrics presented on this study are out of sample evaluations, in other words they are based on estimations on data that the model has not been trained on. These metrics are compared to the $\sqrt{PEHE}$ and ATE of the model when trained on all features. The ATE is also compared to the GT, which is calculated directly from the data. This allows for determining the shift on the estimation when GANITE is subjected to different kinds of confoundedness, and thus the robustness of GANITE to such scenarios. The model used for this study is defined in the repository[1] of the GANITE project. This will be used to train and run the GANs in order to obtain the performance metrics of interest.

$$\sqrt{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((y_i^1 - y_i^0) - (\hat{y}_i^1 - \hat{y}_i^0))^2} \tag{13}$$

[1] https://github.com/jsyoon0823/GANITE

3 Results

The results are broken up by the kind of experiment under which the IHDP, Twins and synthetic datasets are explored in detail. The value obtained when training over all features is denoted AF.

3.1 Single feature removal

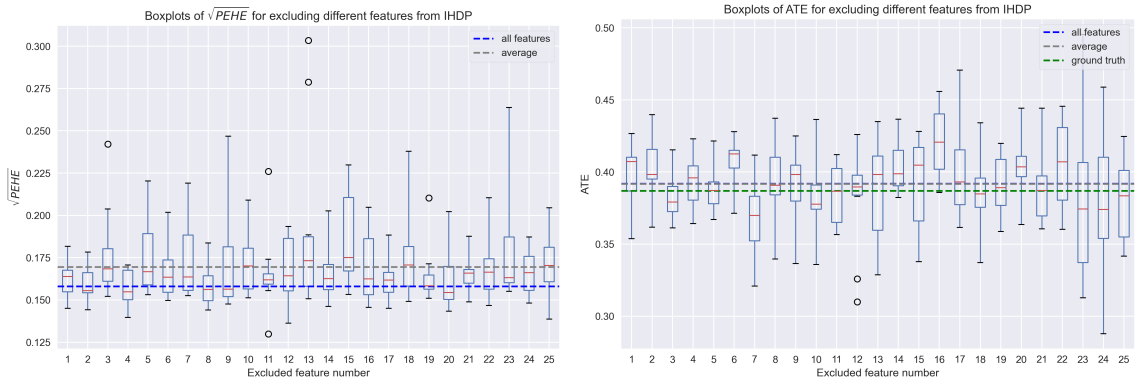
The first experiment performed for the datasets is to remove each feature, one at a time. A total of 250, 300, and 200 trials were conducted for IHDP, Twins, and the synthetic dataset, respectively. Each feature is removed independently from the others and thus the results are expected to be independent as well. For this experiment the results have been plotted through the use of boxplots which illustrate the minimum, 25^{th} percentile, the median, the 75^{th} percentile and the maximum. Any trials that fall more than 1.5 times the interquartile range (75^{th} percentile - 25^{th} percentile) from the extremes are illustrated as points and deemed outliers.

3.1.1 IHDP

Figure 4a shows that on average the $\sqrt{PEHE}$ increases upon the removal of a feature, increasing from 0.158 to 0.170, a 7.60% increase. From the medians of the boxplots upon the removal of most features the error increases. Six out of the 25 features have slightly lower medians although for all their interquartile range overlaps with the error obtained for AF and may thus be due to randomness. Features 13 and 15 have the highest median error which may mean that they have important causal relationships with the treatment and/or the outcome. Feature 13 also has two large outliers which further supports it's perceived importance. The variability observed in the plot is quite high with every box plot overlapping with the average error obtained for AF. Thus no feature removal seems to completely throw off the learning of GANITE although it does seem like the removal of a single feature can already have a significant impact on the estimated ITE. Figure 4b shows the effect of single feature removal on the inferred ATE. It can be seen that the removal of different features can have different impacts on the inferred ATE. Some make the inferred ATE seem higher, most particularly when removing feature 6 and 16, while others make it seem lower such as through the removal of feature 7, 23, and 24. Most features do not have a very large impact with their medians being very close to both the ATE found for AF and the GT. Furthermore, all features have minima and maxima that overlap with the GT. It is interesting to see that when averaging over all of the trials the ATE is practically equivalent to the inferred ATE of AF. This is most likely because when averaging over all of the trials the removal of single features is compensated by the other trials where that feature was present.

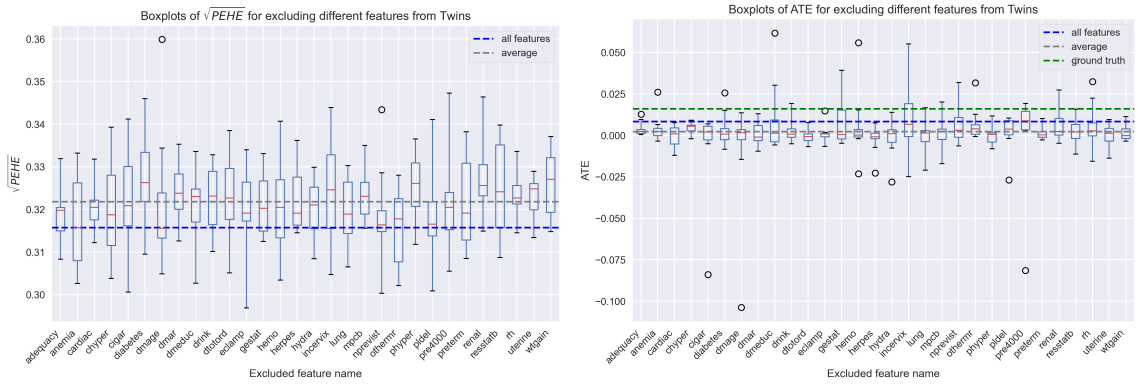
3.1.2 Twins

Similarly to IHDP it can be seen from Figure 5a that upon the removal of single features the $\sqrt{PEHE}$ tends to go up - in this case from 0.316 to 0.322, a 1.90% increase. This error may be lower than for IHDP because one feature is a smaller proportion of the data and thus GANITE is better able to recover. Although this difference is quite small the medians of each boxplot except for one, the "anemia" feature, are above the error of AF. No particular feature removal seems to be more detrimental to the learning of the model. The first thing to notice from Figure 5b is that the average ATE over all trials is further from the GT when compared to the inferred ATE of AF. Most features have relatively small interquartile ranges apart from "dmeduc" and "incervix". Possible reasons for this are explored in section 4. Unlike in IHDP some of the marked outliers are very far from the GT ATE, even flipping in polarity, the extremest being a trial for the "dmage" feature which infers an ATE of -0.104 instead of the GT of 0.0160.



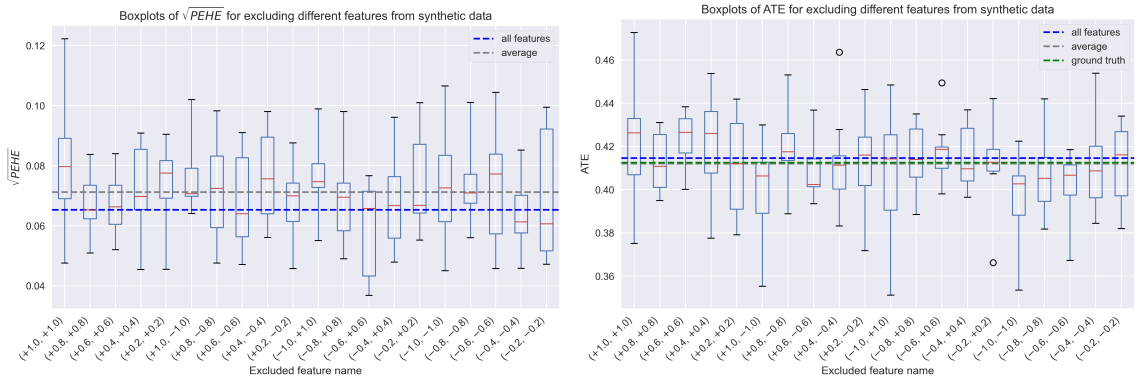
(a) $\sqrt{PEHE}$ (b) ATE with the GT

Figure 4: Boxplots of $\sqrt{PEHE}$ and ATE for IHPD for each removed feature with lines for the average error of AF and over all single feature removal trials



(a) $\sqrt{PEHE}$ (b) ATE with the GT

Figure 5: Boxplots of $\sqrt{PEHE}$ and ATE for Twins for each removed feature with lines for the average error of AF and over all single feature removal trials



(a) $\sqrt{PEHE}$ (b) ATE with the GT

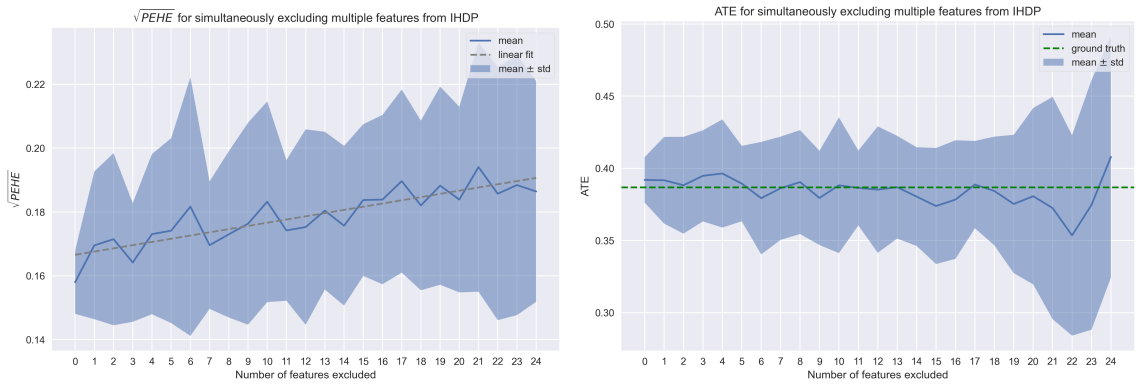
Figure 6: Boxplots of $\sqrt{PEHE}$ and ATE for synthetic data for each removed feature with lines for the average error of AF and over all single feature removal trials

3.1.3 Synthetic data

Unlike for IHDP and Twins the causal effects on the treatment and outcome are known for the features of this dataset. That is why the x-axis in the plots of Figure 6 are ordered primarily by confounder type (+ +, + −, − +, − −) and secondarily by causal effect strength in descending order. In Figure 6a the previously observed relationship of a higher $\sqrt{PEHE}$ is seen once again. For this dataset the error goes up from 0.0654 to 0.0712, a 8.87% increase. All features' boxes either overlap or are entirely above the AF error and thus the fact that the medians of (–0.2, –0.2), (–0.4, –0.4) and (+0.6, –0.6) drop below the line can be explained by randomness. Removing the features with the highest causal effects, namely those with strengths of magnitude 1.0, have higher median errors than the average apart from (+1.0, –1.0). There are no clear patterns for the different kinds of confounders nor for the different strengths of causal effects. The boxplots for the ATE in Figure 6b also show that different features have varying impacts on the inferred ATE. The most evident pattern, if any at all, is that of the (–, –) confounders, where the inferred ATE seems to go down the stronger the causal effect of the removed feature. Unlike in the previous two datasets the average ATE over all trials is actually closer to the GT than to AF, although this difference is very small and when compared to the variability of the results it may very well be an arbitrary result.

3.2 Simultaneous feature removal

In the second category of experiments conducted for the three datasets; subsets of the features are simultaneously removed. A total of 1200, 1460 and 940 trials were conducted for IHDP, Twins, and the synthetic data, respectively. These trials are random samples without replacement and serve as an approximation for the effect of the average over all possible combinations of removing that number of features from the feature space. The results are visualized through a combination of a line plot is used showing the mean for each number of removed features together with a ribbon around it representing the respective standard deviation. The latter is crucial because a single line would hide the high variance present in the results. Furthermore the ribbon provides insight into how this variance changes across experiments. A linear line of best fit, and a the GT are plotted on the $\sqrt{PEHE}$ and ATE plots, respectively.



(a) $\sqrt{PEHE}$ with a linear line of best fit

(b) ATE with the GT

Figure 7: Line plots with shaded bounds representing a ± standard deviation from the mean for $\sqrt{PEHE}$ and ATE for IHDP under simultaneous feature removal

3.2.1 IHDP

Upon removing features simultaneously from IHDP an upwards linear trend of $\sqrt{PEHE}$ in Figure 7a is seen. The fit goes up from 0.167 for AF to 0.191 when training on only one, an increase of 14.4%. The error increases rapidly after removing a single feature although this seems like it may be exaggerated as it does drop down again for the removal of 3 features. Overall there is a fair degree of variance although the trend seems strong and the variability can be explained by the non determinism of GANITE itself. The ribbons increase in width meaning that the spread of the errors increases as more features are removed. More concretely, the standard deviation increases from 0.0099 for removing none to 0.0346 when removing 24 features, an increase of 249%. Moving to Figure 7b the mean line does not seem to have any particular trend at all and generally stays close to the GT. What is striking is that the ribbons become much wider as more features are excluded showing that GANITE is increasingly uncertain of the ATE. The standard deviation increases from 0.0157 for removing none to 0.0835 when removing 24 features, an increase of 432%. This is particularly apparent after more than 19 features are removed at a time, accounting for roughly 80% of the data.

3.2.2 Twins

As more features are removed from the Twins dataset, Figure 8a does not show any clear trend in the $\sqrt{PEHE}$. The linear fit increases from 0.320 to 0.325, an increase of 1.56%. The variability does seem to increase significantly although this pattern is very spiked. The standard deviation increases from 0.00694 for AF to 0.0677 when training on only one, an increase of 876%. This increase is likely a large overestimation as the standard deviation for the trials where all but three features are removed is 0.00978, only 40.9% higher than AF. The results for the ATE, as shown in Figure 8b, has two main similarities to the corresponding $\sqrt{PEHE}$ plot. The first is that the mean line doesn't seem to have much of a trend at all, and the second is that the variance of the standard deviation is very high over the last 5 experiments, from removing [25, 29] features simultaneously. The standard deviation increases from 0.00870 to 0.133, an increase of 1430%, for AF to when trained on only one. Just like for $\sqrt{PEHE}$, the standard deviation is significantly less large for the trials trained on 3 features, 0.0201, which is only 131% larger. One thing to note specifically about this plot is that the ATE is consistently underestimated relative to the GT, something that does not seem to change

(a) $\sqrt{PEHE}$ with a linear line of best fit

(b) ATE with the GT

Figure 8: Line plots with shaded bounds representing a ± standard deviation from the mean for $\sqrt{PEHE}$ and ATE for Twins under simultaneous feature removal

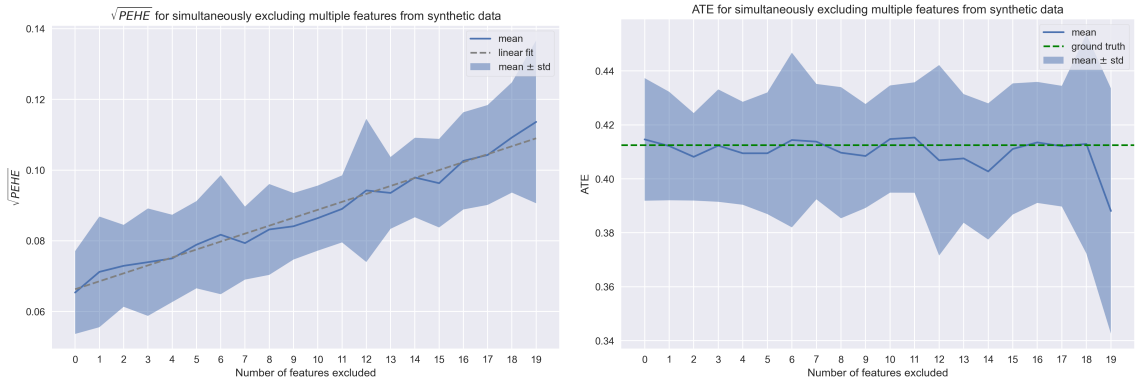
when altering the hidden confounders. This weak trend line for $\sqrt{PEHE}$ and the high variability in standard deviation for both plots is discussed in section 4.

3.2.3 Synthetic data

In Figure 9a the linear trend in $\sqrt{PEHE}$ is very clearly upwards increasing from 0.0663 to 0.109 from removing no features to removing all but one, a 64.4% increase. Furthermore, the ribbons around the mean increase towards the extreme cases of removing more than 16 features. The standard deviation for AF is 0.0117 which increases to 0.0137, 0.0141, 0.01556, and 0.0230 for the removal of 16, 17, 18, and 19 features. The relative increase is 17.1%, 20.5%, 33.0%, 96.6%, respectively. From this it seems like there is a relationship between the variability of the error and the number of features removed but the steep increase for the experiments with 19 removed features is unique, as is frequently the case when testing boundaries. Similarly to IHDP and Twins Figure 9b shows that for the synthetic dataset the removal of features does not cause a drift for the ATE in any particular direction. The standard deviation of the inferred ATE actually stays relatively constant apart from the trials where 18 and 19 features are removed. The change is negligible between AF, 0.0227, and trials where 17 are removed, 0.0224 (less than 1% percentage difference). While the standard deviation spikes for the last two trials with 0.0406 and 0.0455, 78.7% and 100%, respectively. Similarly to the $\sqrt{PEHE}$ plot this is most likely due to the extreme nature of these trials where more than 90% of the original data has been removed.

4 Discussion

From the results of the single feature removal it is clear that hypothesis #2 holds as the $\sqrt{PEHE}$ increased on average for all three datasets - this is particularly apparent for the IHDP and the synthetic dataset. Meaning that upon the removal of a feature GANITE's performance can already suffer. Whether hypotheses #1 holds is much less clear: for IHDP and Twins it is seen that the error increases a lot under the removal of certain features over others but their causal graphs are not known. The "dmeduc" and "incervix" features are not highlighted to be of particular importance in the work of Almond et al. (2005) even though the predictions of their ATE vary a lot upon their removal. The most important confounder is said to be "cigar" although no extreme errors



(a) $\sqrt{PEHE}$ with a linear line of best fit

(b) ATE with the GT

Figure 9: Line plots with shaded bounds representing a ± standard deviation from the mean for $\sqrt{PEHE}$ and ATE for synthetic data under simultaneous feature removal

nor deviations from the inferred ATE are found. For the synthetic dataset the results show that removing features with strong causal effects does generally lead to higher errors and deviations in ATE although more experimentation would need to be performed to confirm this. One of the primary goals of this set of experiments was to determine the effect of removing specific categories of confounders: unfortunately the plots do not show any clear patterns as the variability of the predictions is so high that any observed patterns may be due to noise. The number of trials would need to be increased significantly in order to make strong conclusions for hypotheses #3. The causal effects of the removed features may not have been strong enough to cause significant changes. Given that for the synthetic dataset all features are known confounders it is surprising that there are still features for which the error doesn't seem to increase very much at all, and in certain cases actually falls below the error obtained when training over all of the features. This is almost certainly due to the instability of the underlying GANs of the model.

The simultaneous feature removal experiments show varying results for the trend in $\sqrt{PEHE}$ as more features are removed. For IHDP the trend supports hypothesis #4 but the large standard deviation makes the result uncertain. For the synthetic dataset the trend is very clear and strongly supports the hypothesis. The Twins dataset shows very different results, namely the error does not seem to increase at all even after more than half of the features are removed. The reported $\sqrt{PEHE}$ is in line with the optimal error reported in Yoon et al. (2018) and thus these results are not unique to this study. Theoretically this could occur if the features do not have any confounding effects and thus removing them does not hamper the learning of the ITE. Although what seems more likely is that the model fails to converge to a generalizable solution and that the predictions, although achieving a low error, were poor in the first place. There does not seem to be any particular trend in inferred ATE as more features are removed for any of the datasets, which is inline with hypotheses #5, which may be due to the datasets being made up of confounders with varying causal graphs just like the synthetic dataset and thus upon removing multiple confounders the inferred ATE is not affected. One of the most interesting findings of these experiments is that hypotheses #6 seems to hold for both $\sqrt{PEHE}$ and ATE - this pattern does not seem to be as strong for the synthetic dataset which may be due to the underlying simplicity of the generation process. For Twins the standard deviation varies a lot in the last 5 experiments, this is most likely due to the hyper-parameters no longer being optimal and thus GANITE severely over fitting. This relationship needs to be further explored by fitting a line to the standard deviation itself as looking at the endpoints

exaggerates the differences.

All in all, it has been very difficult to obtain reliable and precise results from GANITE. Even when using the optimal hyper-parameters for IHDP and Twins the $\sqrt{PEHE}$ and ATE vary significantly between trials. This is expected as GANITE is built on top of two GANs which as discussed by Saxena and Cao (2022) are known for their instability. As features are removed the optimal hyper-parameters are likely to change and thus this variability is further increased. Tuning these for each experiment was simply not feasible given the computational power required. The hyper-parameters were kept constant throughout the trials as an attempt to control for these variables although it can be argued that optimizing them for each experiment would have provided more independent results.

5 Conclusion

This work aims to explore the robustness of GANITE to hidden confounders. This was done through two different groups of experiments spanning three datasets: two semi-synthetic and one generated specifically for this study. The robustness is measured through $\sqrt{PEHE}$ relative to the error obtained when training on all of the features and the deviation of the inferred ATE compared to the GT. As hypothesized the $\sqrt{PEHE}$ went up as individual features were removed and when removing multiple features at a time as shown in section 3, although contrasting results were found for the Twins dataset as discussed in section 4. It was hypothesized that the inferred ATE would shift based on the causal graph of the removed confounder although this was not observed from the experiments. As more confounders are removed from the dataset the standard deviation of the $\sqrt{PEHE}$ and inferred ATE goes up implying that it becomes harder for GANITE to converge to the same state. All in all GANITE seems to be robust to small amounts of confoundedness although becomes increasingly unstable as more is removed. Furthermore it seems like hyper-parameter tuning is crucial in order to have precise predictions even when making small adjustments to the dataset.

6 Future Work

The behaviour of GANITE under the relaxation of the confoundedness assumption is explored but this is only one of several assumptions that the Rubin-Neyman model makes - it would be worthwhile to explore the importance of for example the overlap assumption as well. From the section 4 several important future research paths come to light: the hyper-parameters should be tuned for each experiment in order to account for varying optimums as the datasets are adjusted and more trials are required for each experiment in order to reduce the variance of the results. The variability of the standard deviation should also be explored in more depth as trends seem to be present in subsection 3.2 that are not made concrete in this work. It would also be interesting to vary the synthetic confounders in terms of strength and underlying distributions - for example confounders of a uniform or exponential distribution could be explored to see if similar patterns occur. Throughout this work the importance of a feature is quantified through the strenght of it's causal effects. Unfortunately this is not a single number as it requires both the strength on the treatment and the outcome to be well defined. It would be worthwhile to find a single metric that quantifies this, possibly also for confounders with more complex graphs. Further experiments that explore the effect of the meta properties of the datasets on the robustness of GANITE would be of interest as the results found in this paper may differ as the dimensionality of the features changes and/or the number of instances is varied.

7 Responsible Research

In practice it is very hard, if not impossible, to measure all features that have a confounding effect on treatment and outcome. Thus when trying to estimate the effect of a treatment it is typically under a scenario where there is a degree of confoundedness. This leads to a bias in the estimation of the effect. Considering that ITE inference models are of particular interest in fields like medicine it is crucial to know the uncertainty of the estimated effects in order to decide when to rely on them and when not to. This research has shown that this uncertainty is dependent on the causal strength of the hidden confounders and the total number that are removed.

This work was based on semi-synthetic and synthetic datasets. This is because by knowing the counterfactuals the hypotheses can actually be validated. The synthetic dataset was crucial for verifying the hypotheses concerning the causal graphs of the confounders. For example: with no knowledge of the GT ITE there is no way to determine what the actual error is of the model and whether it gets worse when removing confounders. Thus it can not be guaranteed that these findings extend to real world datasets as their possible characteristics are far more diverse than what was explored in the search space of this study. Although this study does consider datasets with the number of instances varying in orders of magnitude, it does not reach the limits of the datasets found in practice. That being said it does seem likely that the $\sqrt{PEHE}$ and variability of GANITE's predictions will increase as the confoundedness is increased. These findings may not hold for larger datasets, especially considering that "the more plentiful data makes it more mysterious and, therefore, harder to model responsibly" (Guo et al., 2021, p. 2).

When working with computationally demanding models a large bottleneck for obtaining meaningful results is computational power. The experiments conducted for this study can be replicated on a high end laptop running over the course of a couple of weeks. If this is not available then replication will prove costly and time consuming. Being able to increase the number of repetitions by an order of magnitude was not feasible although would most likely improve the interpretability of the results. This shows that access to computational power can be a large differentiator for producing impactful research.

The method for generating data, optimizing and testing GANITE, and analyzing the results can all be found on the GitHub repository[2] for this work. In order to replicate the results for each dataset simply run the respective experiments script. In order to visualize the results the plotting scripts should be used. The results obtained will not be identical due to the stochastic nature of the model and because the dataset is split up randomly for each trial but they should be very similar. Increasing the number of repetitions for each experiment will likely result in results with a lower variance.

References

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The Costs of Low Birth Weight. *The Quarterly Jounral of Economics*, page 54.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *NIPS*, page 9.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2021). A survey of learning causality with data. *ACM Computing Surveys*, 53(4):1â37.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[2] https://github.com/vcovo/cse3000

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal Effect Inference with Deep Latent-Variable Models. Number: arXiv:1705.08821 arXiv:1705.08821 [cs, stat].

Richens, J. G., Lee, C. M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. page 15.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Saxena, D. and Cao, J. (2022). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Computing Surveys*, 54(3):1–42.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. Number: arXiv:1606.03976 arXiv:1606.03976 [cs, stat].

van der Schaar, M. and Maxfield, N. (2021). Individualized treatment effect inference.

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*.