Empirical Analysis of Confounding Bias in Feature Representations for Average Treatment Effect Estimation

by

Marco van Veen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 2nd, 2024 at 14:00.

Student number:	5062322	
Project duration:	November 13th, 2023 – July 2nd, 2024	
Thesis committee:	Prof. dr. ir. M. J. T. Reinders,	TU Delft, advisor
	Dr. ir. J. H. Krijthe,	TU Delft, daily supervisor
	PhD candidate R. K. A. Karlsson,	TU Delft, daily co-supervisor
	Dr. L. C. Siebert,	TU Delft, committee member

Empirical Analysis of Confounding Bias in Feature Representations for Average Treatment Effect Estimation

Marco van Veen

Abstract

Causal inference methods are often used for estimating the effects of an action on an outcome using observational data, which is a key task across various fields, such as medicine or economics. A number of methods make use of representation learning to try to obtain more informative feature representations, which are then used for effect estimation. However, such feature representations can introduce confounding bias in the results if they lose confounding information contained within the original features. In this work, we evaluate an existing metric for measuring confounding bias in representations and use this metric to study two representation learning methods in terms of their biases in settings with low overlap between treated and control populations. We show that the metric is a suitable measure for the amount of confounding bias in a representation and representation learning methods that minimise this bias lead to better average treatment effect estimation in our experiments. The code used for the experiments can be found on GitHub.

1 Introduction

Determining the causal effects of actions on outcomes is a crucial task across many fields, such as medicine or economics. In medicine, for example, researchers are often interested in understanding the effect of some newly developed treatment on the recovery of a patient. Or, in economics for instance, a central bank would want to understand what the effect of a new interest rate policy would be.

One method for estimating the causal effect of some treatment in medicine is to perform a randomised controlled trial. Here, patients are randomly assigned either treatment or no treatment in order to create two separate groups, the treated and control groups. Since treatment was assigned randomly, the two groups should have relatively similar characteristics on average if enough patients were sampled. Taking the difference in average outcomes between the groups should then reflect the effect of the given treatment.

However, such randomised trials are not always feasible to perform due to costs or ethical concerns. Therefore, observational datasets, such as anonymised patient records from a hospital, often have to be used instead. In such datasets, treatments are not randomly assigned and often depend on the specific patient characteristics. For example, it might happen that younger patients are assigned treatments more frequently. This can lead to an imbalance in the treated and control groups, as the treated group consists of a relatively larger amount of young patients compared to the control group. Additionally, age likely also has an effect on the outcomes, as younger people tend to be healthier overall. Features which affect both treatment assignment and the outcome, such as age in this case, are also known as confounders. All of these confounding features must be taken into account when estimating causal effects, since unobserved confounders may lead to correlations between treatments and outcomes that can be wrongly interpreted as causal effects. The bias arising due to failing to account for all confounding factors is also known as confounding bias.

To estimate causal effects from observational datasets, treatment effect estimators, such as inverse probability of treatment weighting (IPW), have been developed to try to weigh the samples in order to achieve balance across the covariate distributions, which reflect the two groups' characteristics. A common choice for the weights is the reciprocal of the propensity score introduced by Rosenbaum and Rubin (1983), which is the probability of an individual to receive treatment conditioned on all covariates. However, obtaining weights using estimated propensity scores can lead to instability in the estimators in case of poor overlap between the covariates of the treated and control groups, in which case mostly, or exclusively, samples from only one of the groups are observed in certain regions of the distributions. This causes the propensity scores to have values close to 0 or 1. Instability in the estimators then occurs, because the samples are divided by these propensity scores during the weighing process.

D'Amour and Franks (2021) introduced the idea of deconfounding scores with the goal of improving the stability of the estimators under such poor overlap settings. Deconfounding scores are feature representations that can be used to obtain more stable weights, without introducing additional biases in the estimates. These deconfounding scores represent all functions of the covariates that lead to unbiased estimates of the treatment effects as long as the original covariates are unconfounded. The covariates are considered unconfounded if they contain all information relevant to confounding.

The authors additionally derived a formulation for the confounding bias which could be introduced when using an alternative representation of the original covariates, as using such alternative representations may lead to a loss of confounding information contained within the original features. Using this formulation, the authors managed to analytically derive a family of deconfounding scores for a simplified setting with low overlap. In that specific setting, using some of the obtained deconfounding scores to calculate the weights can lead to significant improvements over the commonly used propensity scores.

Besides weighting-based estimators, deep learning methods which utilise representation learning have also been developed for estimating treatment effects (Shalit et al., 2017; Johansson et al., 2022; Shi et al., 2019; Chernozhukov et al., 2022). These methods use neural networks to obtain feature representations for the original data which are then used for the outcome predictions. In order to improve the treatment effect estimation, some of these methods also incorporate additional constraints for the learned representations to achieve better balance between the treated and control groups within these learned representations compared to the original features.

However, the feature representations learnt by such deep learning models might also lead to a loss of confounding information contained within the original confounders (Melnychuk et al., 2023). While some methods may include additional balancing constraints on the feature representations, they do not explicitly enforce that no confounding bias is introduced by using the representations. So, this may lead to the resulting treatment effect estimates being affected by confounding bias due to the learned representations.

Since the previous work on deconfounding scores, which minimise confounding bias, showed that several of these scores can lead to better treatment effect estimation for weighting methods, especially in low overlap settings, understanding whether this effect also carries over to representation learning methods can provide valuable insights into whether prioritising confounding bias minimisation may be a desirable objective for representation learning.

Therefore, the goal of this paper is to answer two main questions: (1) Is the confounding bias formulation from D'Amour and Franks (2021) a suitable metric for the confounding bias introduced by feature representations? (2) Do feature representations which lead to lower average treatment effect bias and variance in low overlap settings, also exhibit a lower amount of measured confounding bias within the representations? To address these questions, we take an empirical approach by constructing specific simulation experiments to understand how accurately the confounding bias can be measured using feature representations and to what extent measures of confounding bias can be used to explain the performance of representation learning methods.

The rest of the paper is structured as follows. Section 2 will show an overview of the related works. Next, Section 3 introduces the problem of estimating average treatment effects, including two estimators which can be used for this, and finishes with the main problem setting considered in this work. Section 4 will then explain how confounding bias can be measured within representations and give an overview of the two representation learning models considered in this work. After this, Section 5 presents the setup for the experiments and Section 6 will show the obtained results. Then, Section 7 provides a discussion of the results. Finally, Section 8 summarises the resulting conclusions.

2 Related Work

Deconfounding scores are feature representations of the covariates that can be used to obtain weights for weighing estimators, without losing any information related to confounding, which could otherwise introduce confounding bias in the estimates (D'Amour and Franks, 2021). They belong to a rich literature on balancing scores that try to find weights, such that the covariate distributions are more balanced among the treated and control samples. These works generally find balancing weights through some optimisation problem where constraints are added for the weights such that they match the moments of the distributions between the groups. For example, Athey et al. (2018) find weights which approximately match the means of the distributions in order to weigh the outcome regression residuals. Hainmueller (2012), on the other hand, introduced the entropy balancing method, which explicitly maximises the entropy of the weights such that moments of the distributions are exactly matched. Finally, Imai and Ratkovic (2014) uses the balancing property of the IPW estimator to obtain the weights which balance the covariate distributions.

Only Clivio et al. (2023) seem to build further upon the concept of deconfounding scores. They make use of approximate deconfounding scores, which are deconfounding scores that allow for some fixed amount of confounding bias to be introduced due to the representation. The goal is then to learn feature representations, using e.g. a RieszNet (Chernozhukov et al., 2022), which bound the amount of confounding bias allowed. These are then used to obtain the optimal weights which balance the feature distributions according to some probability distance.

While our work also makes use of the notion of deconfounding scores, they are used for different purposes compared to Clivio et al. (2023). Instead of using deconfounding scores to learn representations which allow for a certain level of confounding bias, here they are used to measure the confounding bias in existing representation learning methods in order to better understand whether minimising the confounding bias in learned representations is actually a desirable objective. Furthermore, understanding the role of the confounding bias specifically in low overlap settings is the main interest of this work, while this setting was not further explored by Clivio et al. (2023).

In this study, we are considering confounding bias in representations learned using neural-network based methods for causal effect estimation. One of the most popular neural network-based methods is the TARNet introduced by Shalit et al. (2017), which contains shared layers to learn a feature representation for the original features and uses two different output heads for predicting the outcomes for the treated and control groups. In order to improve the performance under heavy imbalance between the groups, they also add a balancing objective in the form of an Integral Probability Metric (IPM) to minimise the distance between the treated and control samples in the learnt feature representation (Müller, 1997). The TARNet with the IPM objective is known as the Counterfactual

Regression (CFR). Another extension to the TARNet, called DragonNet, was proposed by Shi et al. (2019). They add an additional treatment prediction head in order to ensure that the feature representation only captures confounding information from the original covariates, while discarding all other information which is not relevant for confounding. For example, information from prognostic features, which only affect the outcome, but not treatment, could be ignored when learning a representation due to the treatment prediction objective, if the representation does not have the capacity to retain all information from the original covariates. They showed that focusing on only the confounding information in the representation layers improves the performance for smaller sample sizes compared to the TARNet. A variant of the DragonNet, called RieszNet, was recently proposed by Chernozhukov et al. (2022). The model can be seen as a generalisation of the DragonNet and manages to outperform the DragonNet when estimating average treatment effects.

A recent work which also deals with confounding bias in representation learning methods is Melnychuk et al. (2023). They provide a framework which allows for estimating lower and upper bounds on the bias in the final treatment effect estimates which comes from introduced confounding bias due to the learned representations. In our work, however, the confounding bias is measured directly from the learned representations and compared to the overall bias in the final treatment effect estimates, which may contain biases from other sources such as lack of overlap, which is the specific setting considered here. Additionally, the estimand of interest in Melnychuk et al. (2023) is the conditional average treatment effect, which looks at treatment effects within subgroups of the population, while this work focuses on population-level average treatment effects.

3 Problem Formulation

Before we can understand whether the confounding bias metric from D'Amour and Franks (2021) is a suitable measure and how it can be used for measuring confounding bias in feature representations, we have to introduce the specific problem setting in this work. To do this, the treatment effect of interest in this work, namely the Average Treatment Effect (ATE), is first defined. After this, two estimators for the ATE are presented, including the assumptions that are required for correct causal effect estimation from observational data. Finally, ATE estimation using feature representations is discussed, along with the potential issues arising from using such representations, which leads to the main problem of interest in this work.

3.1 Average Treatment Effects

The effect we are interested in here is the Average Treatment Effect (ATE) of a binary treatment using observational data. Such a dataset consists of n samples with features X, binary treatments T, and outcomes Y which are independently and identically distributed as $(X_i, T_i, Y_i) \sim P$ for $i = 1, ..., n$. The features X can affect the treatment assigned to each sample, the outcome, or both. Here, capitalised letters represent random variables, while lower case letters, which appear later in this work, represent actual values.

In order to estimate average treatment effects, the Rubin-Neyman potential outcomes framework is used (Splawa-Neyman et al., 1990; Rubin, 1974). Here, it is assumed that each sample in the dataset has two potential outcomes $Y(1)$ and $Y(0)$ which represent the outcomes if the sample had or had not received treatment. Furthermore, the stable unit treatment value (SUTVA) assumption is commonly made, which essentially states that if a sample x receives treatment $(t = 1)$, then its observed outcome Y equals $Y(1)$. Conversely, if it received no treatment, then its observed outcome Y equals $Y(0)$. So, the observed outcome for each sample matches the correct potential outcome according to the assigned treatment.

Now, the ATE can be formulated:

$$\tau^{ATE} = E[Y(1) - Y(0)]$$

which is the difference in the expected outcome in the population, if every individual received treatment, and no individual received treatment.

3.2 Estimators For Average Treatment Effects

The problem with estimating the ATE is the fact that only one of the potential outcomes can be observed for each sample. Therefore, the strong ignorability conditions are used in order to make the ATE identifiable from observed data (Rosenbaum and Rubin, 1983). These conditions consist of two separate assumptions. The first is the unconfoundedness assumption which states $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$. This means that, conditioned on the observed features X, the treatment assignment is essentially random. The second is the overlap assumption $0 < P(T = 1|X) < 1$, where $e(X) = P(T = 1|X)$ is also commonly known as the propensity score. This ensures that each individual with features X has some non-negative probability to either receive treatment or no treatment. Under these assumptions, the ATE can be estimated from the data by conditioning on X and T:

$$\tau^{ATE}_{cond} = E[E[Y|X, T = 1] - E[Y|X, T = 0]]$$

By defining two functions for the conditional expectations as $m_t(X) = E[Y|X, T = t]$ for $t = 0, 1$, it is possible to estimate the ATE using the parametric G-formula (Robins, 1986). The functions m_0 and m_1 can, for instance, be estimated using regression and be used to estimate the ATE by averaging the predicted the outcomes under treatment and no treatment over all samples, and taking the difference to obtain the treatment effect.

A commonly used, non-parametric method for estimating the ATE is the Inverse Probability of treatment Weighting (IPW) estimator. The IPW reweighs the observed outcome of each sample by dividing the outcomes with the propensity scores. Rosenbaum and Rubin (1983) showed that using the propensity scores $e(X)$ as weights allows for unbiased treatment effect estimation as long as the strong ignorability conditions hold for the features X. The IPW then takes the following form:

$$\tau^{ATE}_{IPW} = E\left[\left(\frac{T}{e(X)} - \frac{1-T}{1-e(X)}\right) Y\right]$$

From the formulation above, it is clear that the IPW estimator can be heavily affected by poor overlap through the division by the propensity scores $e(X)$. In the extreme case of structural overlap violations, where $e(X)$ can attain values of 0 or 1, the estimator breaks down completely. This is also the case for the first estimand τ^{ATE}_{cond} when estimated non-parametrically.

3.3 Estimation Using Feature Representations

While the previous estimators directly make use of the features X, many works are actually interested in finding feature representations of the original features to use for effect estimation instead. Alternative feature representations may, for instance, help with the generalisation capabilities when using models to predict the counterfactual outcomes of the treated and control samples. So, they try to learn feature representation $d(X)$ which map the original features X to another space, which may have either a higher or lower dimensionality compared to the original feature space. Additionally, different objectives can be used for learning these representations, such as trying to enforce more balance between the treated and control groups within the feature space. The representations $d(X)$ are then used to estimate the ATE using $\tau_d^{ATE} = E[E[Y|d(X), T = 1] - E[Y|d(X), T = 0]]$. They can also be used to obtain weights for the IPW estimator τ_{IPW}^{ATE} by estimating the propensity scores using the feature representations $d(X)$ instead of the original features X.

However, such feature representations may lose confounding information contained within the original features X. Losing such confounding information causes the unconfoundedness assumption to be violated, so then $(Y(0), Y(1)) \not\perp\!\!\!\perp T \mid d(X)$. This violation leads to the introduction of confounding bias in the ATE estimates, which is defined as $\tau_d^{ATE} - \tau^{ATE}$.

This confounding bias introduced by feature representations directly leads to the main problems investigated in this work. First, we are interested in how to measure the amount of confounding bias introduced by using a feature representation $d(X)$. Then, the effect of different methods for learning representations $d(X)$ on the introduced confounding bias and other biases in the ATE is investigated. This will be done specifically for the case where there is low overlap between the treated and control distributions, which will help us to understand whether focusing on confounding bias minimisation when learning representations could be more desirable in these challenging settings compared to, for example, focusing on learning a more balanced representation instead.

4 Methodology

This section will provide an overview of the method that will allow us to measure confounding bias in feature representations and two different methods for learning feature representations. First, a formulation for confounding bias is presented, which can be used to measure the amount of confounding bias introduced by using a feature representation. Then, two representation learning methods that have two different objectives for learning the representations will be shown. These will be used to understand the effect of different representation learning objectives on the amount of introduced confounding bias, and how they affect the overall bias from different sources in the resulting treatment effect estimates.

4.1 Measuring Confounding Bias

As previously stated, using a feature representation $d(X)$ can potentially introduce confounding bias in the ATE estimates if the representation leads to a loss of confounding information. D'Amour and Franks (2021) derived a formulation which can measure the amount of confounding bias introduced by using a feature representation:

$$\tau_d^{ATE} - \tau^{ATE} = E\left[\frac{Cov(Y(1), T|d(X))}{e_d(X)} + \frac{Cov(Y(0), T|d(X))}{1 - e_d(X)}\right]$$

where $e_d(X) = P(T = 1|d(X))$ is known as the reduced propensity score. Feature representations $d(X)$ which do not introduce any confounding bias, i.e., set the above equation to 0, are called deconfounding scores by D'Amour and Franks (2021).

It has to be noted that the formulation above differs slightly from the one presented in the original paper, since they used $\tau^{ATE} - \tau_d^{ATE}$ for the left hand side of the equation, while keeping the same expectation on the right. That, however, appears to be a small mistake in the formulation, as the expectation should be negative in that case. This is shown by providing an overview of the proof for the confounding bias formulation in Appendix A, which is based on the one provided in D'Amour and Franks (2021).

This bias formulation will lead to a strategy for estimating how much confounding bias is introduced in the ATE estimates, without actually using the obtained ATE estimates themselves, but only the feature representations $d(X)$. To do this, the covariances between the potential outcomes and treatments must be measured for each level of $d(X)$.

There are, however, two problems with this approach to measure the confounding bias.

First, only one of the potential outcomes is generally observed for each sample, while both are required in order to measure the confounding bias. So, the true bias can only be measured from $d(X)$ in synthetic data experiments where both potential outcomes are known. It is, however, possible to estimate the confounding bias by instead using models to estimate the propensity scores and the outcomes under treatment and no treatment, which can be done by learning the functions for the conditional expectations $m_t(X) = E[Y|X, T = t]$ for $t = 0, 1$. Under the assumption that the strong ignorability conditions hold for the original X, the confounding bias can then be estimated from the observed data using $E\left[\frac{Cov(m_1(X),e(X)|d(X))}{e_d(X)} + \frac{Cov(m_0(X),e(X)|d(X))}{1-e_d(X)}\right]$ (D'Amour and Franks, 2021).

Second, since these covariances are conditioned on $d(X)$, they must be estimated at each level of $d(X)$, which is generally continuous and possibly high-dimensional. One strategy for dealing with this issue, which will be used in this work, is to divide the space of the feature representation into a number of small bins and estimate the covariance for each bin. This requires a sufficient number of samples per bin in order to accurately estimate the covariances. This can especially be problematic if $d(X)$ is high dimensional, as the number of bins then increases exponentially.

4.2 TARNet

The TARNet is a neural network-based model for estimating treatment effects introduced by Shalit et al. (2017). An overview of the architecture can be seen in Figure 1a.

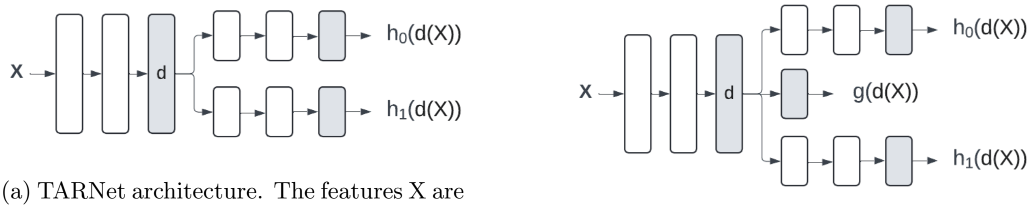
Instead of fitting two separate models on the treated and control groups to estimate the potential outcomes $Y(0)$ and $Y(1)$, it tries to combine feature information from both groups by passing all samples through a number of shared layers in order to learn a more informative feature representation $d(X)$. After these shared layers, the samples are divided into two separate outcome heads depending on the samples' assigned treatments in order to learn two models $h_0(d(X))$ and $h_1(d(X))$ for the potential outcomes $Y(0)$ and $Y(1)$. The objective function to be minimised then consists of some loss function, in this case the mean squared error (MSE), between the predicted and observed outcomes:

$$\mathcal{O}_{TARNet}(h_0, h_1, d) = \frac{1}{n}\sum_{i=1}^{n}(h_{t_i}(d(x_i)) - y_i)^2$$

The two obtained outcome models can then be used to estimate either the conditional average treatment effect (CATE) or the ATE, which is the estimand of interest in this work. The ATE can be estimated using the two learned outcome models through $\hat{\tau}^{ATE} = \frac{1}{n}\sum_{i=1}^{n}(h_1(d(x_i)) - h_0(d(x_i)))$.

4.3 Counterfactual Regression

An extension to the basic TARNet, which adds a balancing objective, was also proposed by Shalit et al. (2017) and is known as Counterfactual Regression (CFR). This model additionally minimises an Integral Probability Metric (IPM) (Müller, 1997) between the treated and control distributions with the goal of learning a feature representation $d(X)$ that is more balanced compared to the original features X. This representation $d(X)$ with smaller distances between the treated and control features should then lead to better generalisation performance when predicting counterfactual outcomes between the treated and control samples.



(a) TARNet architecture. The features X are put through a number of shared layers and then go through one of two separate output heads depending on the assigned treatment per sample. The outcomes $h_1(d(X))$ and $h_0(d(X))$ represent the predicted outcomes under treatment or no treatment.

(b) DragonNet architecture. The architecture is similar to the TARNet, but it includes an additional propensity head $g(d(X))$ and estimates the probability of treatment for each sample based on its learned feature representation.

Figure 1: Architectures for the TARNet and DragonNet.

Shalit et al. (2017) proposed two different IPMs which could be used, namely the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Wasserstein distances (Villani et al., 2009). Here, only the Wasserstein distance is considered, as this was shown to perform better than the MMD in their experiments. As it can be expensive to repeatedly compute the Wasserstein distance due to a linear program which has to be solved every time, the Wasserstein distance can be approximated using Sinkhorn distances (Cuturi, 2013), which may be minimised instead. An overview of how the Wasserstein distance is approximated in this case can be found in Shalit et al. (2017). The algorithm contains a parameter λ that has to be set, which scales the calculated distances between the treated and control samples in order to assign higher or lower weight to more distant treated and control samples.

The objective function of the CFR to be minimised is then:

$$\mathcal{O}_{CFR}(h_0, h_1, d) = \frac{1}{n}\sum_{i=1}^{n}(h_{t_i}(d(x_i)) - y_i)^2 + \alpha \cdot \text{IPM}_{Wass}(\{d(x_i)\}_{i:t_i=0}, \{d(x_i)\}_{i:t_i=1})$$

where IPM_{Wass} represents the Wasserstein distance between the representations of the control samples $\{d(x_i)\}_{i:t_i=0}$ and the treated samples $\{d(x_i)\}_{i:t_i=1}$. The strength of the IPM term can be modified by changing the value of α, where setting $\alpha = 0$ reduces CFR to the basic TARNet.

4.4 DragonNet

Another extension to the basic TARNet, called DragonNet, was proposed by Shi et al. (2019) and introduces an additional propensity head $g(d(X))$ for predicting treatment assignment. The architecture of the DragonNet is shown in Figure 1b.

The additional propensity head is added to the model in order to force the representation $d(X)$ to only keep information relevant for confounding, which could, for example, be beneficial if the representation lacks the capacity to preserve all information from the original features X. The reasoning behind this is that other information, such as from prognostic features that only affect the outcome, is not relevant for correct causal effect estimation, since only confounders have to be taken into account. So, it can be seen as additional noise when adjusting for confounders when estimating the treatment effects, which may especially degrade estimation performance in smaller samples.

With the additional propensity head, the objective function for the DragonNet is:

$$\mathcal{O}_{DragonNet}(h_0, h_1, g, d) = \frac{1}{n}\sum_{i=1}^{n}\left[(h_{t_i}(d(x_i)) - y_i)^2 + \alpha \cdot \text{CrossEntropy}(g(d(x_i)), t_i)\right]$$

where α determines the strength of the treatment prediction objective and setting $\alpha = 0$ returns the basic TARNet.

5 Experimental Setup

To study whether the covariance metric is a good measure of confounding bias, and to investigate the effects of the two different representation learning approaches on different sources of causal bias, we need to construct a number of simulation scenarios that have various levels of confounding bias, overlap, and complexity. First, we will describe a setting that contains a hidden confounder, which will allow us to evaluate how well the covariance metric is able to capture this introduced confounding bias. Next, the setting used for the different representation learning methods will be presented to understand their effects on the biases. Then, we will describe how additional prognostic and instrumental variables are added to the previous setting, which allows us to assess how different non-confounding features affect the two different models. Since the covariance metric is formulated in terms of potential outcomes, which are not known in practice, we also show a setting where we will evaluate how well the covariance metric can be estimated using predicted outcomes instead. Finally, the implementation details for the different models and covariance metric are given.

5.1 Evaluating the Covariance Metric

To be able to evaluate how good the covariance metric is able to measure the amount of confounding bias in a representation, we construct a simple linear setting which contains a hidden confounder. In this simple linear setting, the only source of bias is the hidden confounder and the amount of confounding bias is reflected by the amount of bias in the final ATE estimates, which are calculated as $\hat{\tau} - \tau$, where $\hat{\tau}$ is the estimated ATE and τ is the true ATE. In these experiments, two linear regression models are fitted on the observed confounders to predict the outcomes under treatment and no treatment. These models can then be used to estimate the ATE by taking the difference between the average predicted outcomes under treatment and no treatment. By increasing the strength of the hidden confounder to increase the amount of expected confounding bias, we can evaluate how well the metric is able to measure confounding bias by comparing the bias measured by the covariance metric against the bias in the final ATE estimates.

For this and all other experiments, we have $T \sim Ber(0.5)$, so around half of the population is expected to be treated on average. The confounders are generated according to $X_i \mid T = t \sim N(\mu_t, 1)$ for $t = 0, 1$, where $\mu_0 = 0$ and $\mu_1 = 1$. The hidden confounder U is generated in the same way. This leads to a scenario that closely resembles a situation with structural overlap violations due to the very small probabilities of observing samples in the tails of the distributions, which is visualised in Figure 2. Finally, the outcome model used is $Y(t) = 0.75X_1 + X_2 + \beta U + 2t + N(0, 0.5)$, where X_1 and X_2 are the observed confounders and U is the hidden confounder with varying strength β.

A different setting with non-linear effect modification, where the treatment effect also depends on the value of some feature, is also considered to understand the effect of concurrent misspecification bias. To do this, we now consider an outcome model of the form $Y(t) = 0.75X_1 + X_2 + \beta U + t(2 + 0.5X_1^2) + N(0, 0.5)$.

In these two scenarios, we do not yet learn feature representations. Instead, the two observed confounders X_1 and X_2 serve as a feature representation that forgets all confounding information contained within the hidden confounder U. So, in this setting, we have a feature representation $d([X_1, X_2, U]) = [X_1, X_2]$.

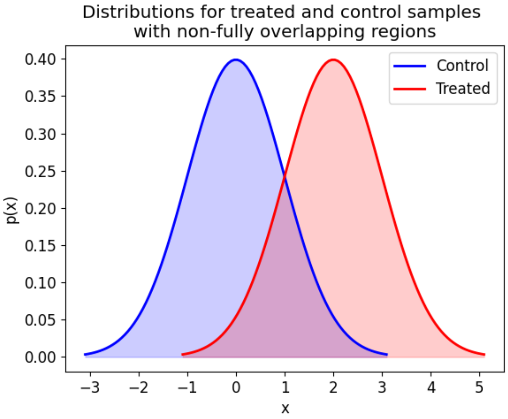


Figure 2: Graph showing the distributions of treated and control samples. The confounders follow standard normal distributions, but the distributions for the treated samples are shifted to the right by varying amounts to decrease the amount of overlap during the experiments.

5.2 Biases from Representation Learning Methods

For these experiments, we will compare the CFR and DragonNet models in a more complex setting to understand the effects of their two different objectives on the confounding bias in their learned representations, and the total bias in the ATE estimates, under decreasing levels of overlap. Here, the covariance metric will be used to measure the confounding bias in the representations. Additionally, the learned representations will also be used as inputs for an IPW estimator in order to understand whether the different objectives of the models can lead to more stable weights in low overlap settings by looking at their ATE biases. This will be done by estimating the propensity scores using the feature representations instead of the original features. A logistic regression is used to fit a model for the propensity scores $e_d(X) = P(T = 1|d(X))$, which will be used as the weights.

The confounders in this scenario are again distributed as $X_i \mid T = t \sim N(\mu_t, 1)$ for $t = 0, 1$. However, now $\mu_0 = 0$ and we let μ_1 vary from 0 to 2, where larger values lead to a decrease in overlap between the treated and control groups. The outcome model is now more complicated due to non-linear terms and has the form $Y(t) = X_1 + 0.25X_2^2 + 0.5X_3^2 + 0.3X_4^3 + 2t + N(0, 0.5)$.

A scenario with non-linear effect modification is also used. Here, the outcomes are generated by $Y(t) = 0.75X_1 + X_2 + \beta U + t(2 + 0.5X_1^2 + X_2) + N(0, 0.5)$.

5.3 Effect of Additional Prognostic and Instrumental Variables

Since the above experiments only consider confounding variables, we will also evaluate the effects of additional, non-confounding variables on the different representation learning models under decreasing overlap. Two types of non-confounding variables are considered, namely prognostic and instrumental variables. Prognostic variables only affect the outcome and do not affect the treatment assignment, while instrumental variables only affect the treatment assignment, but not the outcomes. Two experiments are performed where either a prognostic or an instrumental variable is added to the previous setting, which will allow us to understand the effect both types of variables have on the measured biases under decreasing overlap.

The confounders X_i are again generated as $X_i \mid T = t \sim N(\mu_t, 1)$ for $t = 0, 1$, with $\mu_0 = 0$ and μ_1 varying from 0 to 2. The additional instrumental variable (IV) is also generated as $X_{IV} \mid T = t \sim N(\mu_t, 1)$ for $t = 0, 1$, where the $\mu_0 = 0$ and μ_1 also changes from 0 to 2, depending on the overlap setting. The outcome function remains unchanged, since this variable does not affect the outcome, and is again defined as $Y(t) = X_1 + 0.25X_2^2 + 0.5X_3^2 + 0.3X_4^3 + 2t + N(0, 0.5)$.

The prognostic variable is generated as $X_{prog} \sim N(0, 1)$, such that it has the same distribution for both treated and control groups. The outcome function in this case is $Y(t) = X_1 + 0.25X_2^2 + 0.5X_3^2 + 0.3X_4^3 + 0.5X_{prog}^2 + 2t + N(0, 0.5)$.

5.4 Estimating the Covariances from Observed Data

The covariance metric is formulated in terms of potential outcomes, which can be available in synthetic datasets, but never in real-world settings. Therefore, we will also try to estimate the covariance metric using only the data that can be observed and compare the results to the covariance metric which is estimated using potential outcomes. As described in subsection 4.1, the potential outcomes are replaced by predictions from linear regression models for the treated and control outcomes, and the treatments are replaced by the estimated propensity scores using logistic regressions. Additionally, these models must be fitted on the original unconfounded features.

The setting for this experiment will be the same as the setting with the hidden confounder from before (subsection 5.1), where we have a representation $d([X_1, X_2, U]) = [X_1, X_2]$. Since the models for the outcomes and propensity scores must be learnt using all confounders to correctly estimate the conditional covariances, we fit these models on both the observed confounders X_1 and X_2, and the hidden confounder U.

5.5 Implementation of Estimators

The CFR and DragonNet models used during the experiments consist of 2 shared layers for learning the representations and 1 layer for each of the two output and single propensity heads. As described in Shalit et al. (2017), the layers within the shared representation and outcome blocks are connected using the exponential linear unit activation functions, and batch-normalisation is applied on the outputs of the shared representation. The hidden dimensions are set to 5 and the dimensions of the shared feature representations is set to 2. For approximating the Wasserstein distances when using the IPM, λ is set to 1 for the algorithm which approximates the Sinkhorn distances.

The train and test sets both consist of 10,000 samples for all experiments. If a different test set size is used, it will be clearly specified when presenting the results. A batch size of 512 is used during training along with the Adam optimizer. The number of epochs is set to 25 per run and all results are obtained over 25 runs. The learning rate is set to 0.1, as the IPM often did not seem to converge within the number of epochs when using a lower learning rate.

In order to estimate the covariance term for calculating the confounding bias when dealing with continuous variables, the outputs from the feature representations have to be binned in order to estimate the covariance at each level of the representations. The overall confounding bias is then obtained by averaging over all the non-empty bins. Since the values of the feature representations are not bounded, the obtained representations are first standardised to center them around the origin and all values between -2.5 and 2.5 for each dimensions are considered, since this contains the vast majority of the samples. The number of bins along each dimension of the representation is set to 40 to allow for relatively fine-grained bins. A test set of size 10000 is used to allow for a sufficient number of samples in the bins for estimating the covariances. Additionally, since the confounding bias formulation uses the covariance between the potential outcomes and treatment assignments, the true potential outcomes are included for the test samples.

6 Results

The results from the experiments are discussed in the following sections. First, the results for the case with hidden confounding will be shown in order to evaluate how well the covariance metric can measure confounding bias in representations. Next, results obtained from using the CFR and DragonNet models are presented, which will allow us to compare the effects these two different representation learning approaches have on the measured biases. To understand how different, non-confounding features could affect the models, the following section will show the results for the CFR and DragonNet in the settings with an additional prognostic or instrumental variable. The results up until this point make use of the potential outcomes to estimate the covariances. However, since these are not available in real-world datasets, some results will also be provided in the last section where we try to estimate the covariance term using only observed data instead.

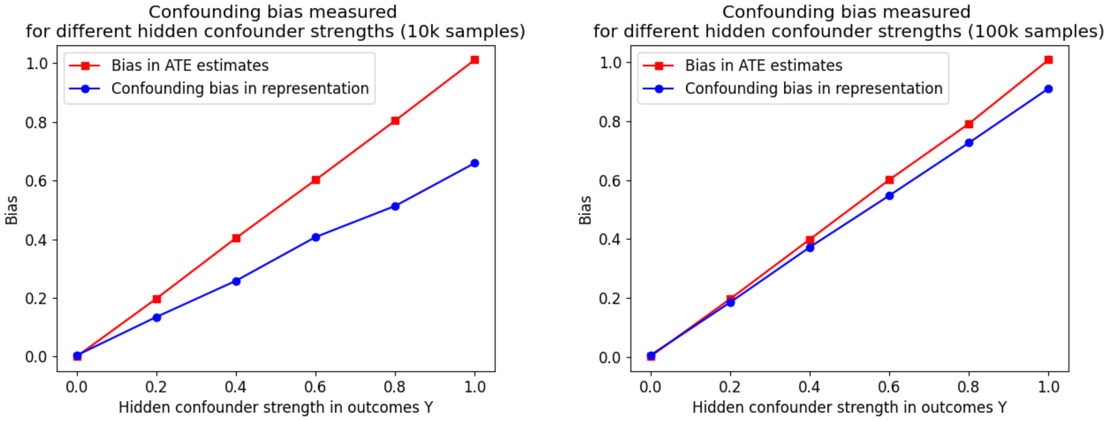


Figure 3: Graphs showing the confounding bias measured in the observed confounders and the bias in the resulting ATE estimate while increasing the weight of the hidden confounder in the outcome model. In this specific setting, the bias in the ATE is solely due to the unobserved confounding. The two graphs present the results when **10,000 (left)** or **100,000 (right)** samples to evaluate the covariance-based measure for confounding bias. Here, it can be seen that as the strength of the confounder increases, meaning a larger amount of confounding bias introduced, a larger difference is observed between the true and measured confounding bias. This difference decreases as the sample size increases.

6.1 Evaluating the Covariance Metric

In order to understand whether the covariance between potential outcomes and treatment assignments is a suitable measure for the confounding bias, we consider a simple linear setting with 2 observed confounders and one hidden confounder, whose strength on the outcomes is gradually increased. This increase in strength of the hidden confounder leads to more confounding bias being introduced. For estimating the ATE, two linear regression models are fitted to the treated and control samples to predict the counterfactual outcomes. We do not make use of the representation learning methods in this experiment, because simple linear regression models can accurately predict the ATE in this linear setting and we are considering a manually constructed feature representation from the observed features.

In this specific setting, we have a feature representation $d([X_1, X_2, U]) = [X_1, X_2]$, which loses the confounding information from the hidden confounder U. So, the confounding bias measured in this "feature representation" can be compared to the bias measured in the ATE estimates to test how well the covariance-based measure can estimate the expected amount of confounding bias as observed in the ATE estimates.

In the case with no unmeasured confounder, the well-specified regression model should be able to accurately predict the average treatment effect due to the simple linear outcome function that allows for perfect extrapolation across non-overlapping regions of the treated and control distributions. So, it is expected that the effect of introducing an unmeasured confounder should be reflected by an increase in the covariance-based metric for measuring confounding bias, especially as the strength of the hidden confounder increases. Additionally, this should be reflected in the resulting ATE bias, as this unmeasured confounder should be the only source of bias in this simple setting.

The results for the setting with unmeasured confounding when using 10,000 and 100,000 samples for estimating the bias are shown in Figure 3. It can be seen that the unmeasured confounder indeed leads to a confounding bias being measured in the two observed confounders which increases as the strength of the hidden confounder on the outcome increases. However, a gap between the expected confounding bias in the ATE and the measured confounding bias can be observed which increases as the amount of confounding bias added through the hidden confounder increases. This difference between the measured and expected bias appears to decrease as a significantly larger number of samples is used. The number of samples required likely depends on a number of factors, such as the amount of confounding bias, specific outcome function used, and binning procedure for estimating the covariances.

This experiment shows that the covariance metric is able to measure confounding bias introduced by using a feature representation which loses confounding information. However, the number of samples required to accurately estimate the bias seems to significantly increases with the amount of confounding bias that is being introduced. We have also considered a setting which contains other concurrent biases, specifically bias through model misspecification. These results can be found in Appendix B, which shows that in certain cases, such as quadratic effect modification in this scenario, different biases may cancel out each other's effects.

6.2 Biases from Representation Learning Methods

In these experiments the more challenging, non-linear outcome model is used for the CFR and DragonNet models. The goal in this setting is to understand the effect of their two different objectives on the confounding bias in the learned representations and the total bias in the resulting ATE estimates in low overlap settings. This will be done by first examining the results for both the CFR and DragonNet models in order to see whether there are differences in their measured biases, and how large these differences are. Next, the learned feature representations from both models are also used as inputs for an IPW estimator instead of using them in the outcome models. The goal of this is to see whether representations with certain characteristics, such as lower measured confounding bias or more balance, lead to differences in IPW performance in this difficult setting with overlap violations.

6.2.1 Estimation using CFR and DragonNet

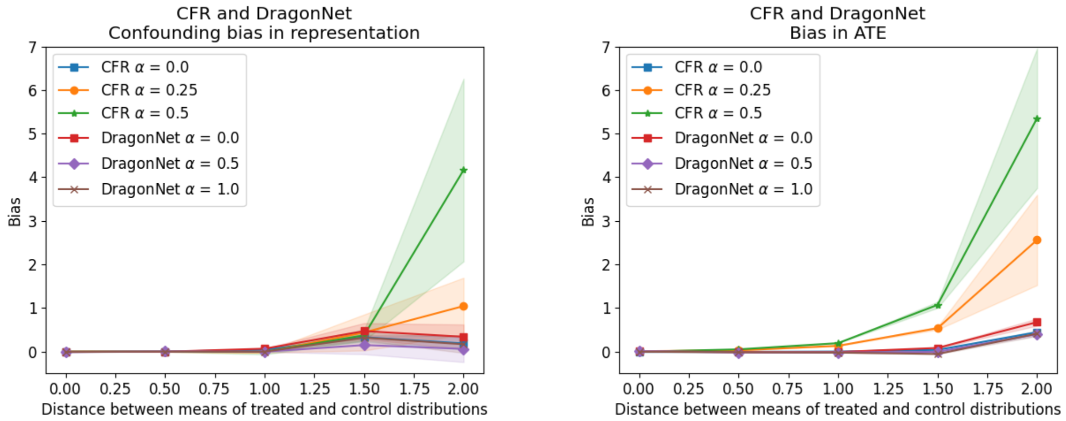


Figure 4: Graphs showing the biases for the different CFR and DragonNet models under decreasing levels of overlap. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates.The different lines show the values of α for each model which correspond to the weights of the IPM or treatment prediction terms in the objective functions for CFR and DragonNet, respectively. These graphs show how the balancing objective in CFR adds a significant amount of confounding bias in the representations in order to perform the balancing. When looking at the ATE biases, it appears that this additional balancing does not manage to lead to better effect estimates, indicating that representations with lower confounding bias appear to be more beneficial.

Since the goal of the IPM in CFR is to achieve more balance between the treated and control groups through the feature representation, this means that the learned representation may lose some information regarding the outcome prediction which is present in the original covariates in favour of better balance. Since retaining treatment assignment information within the representation is also not one of the objectives of the CFR, it is expected that this loss of outcome information might lead to some confounding bias being introduced in favour of more balancing. However, this should still lead to better ATE estimates when dealing with imbalanced groups, as moving the treated and control samples closer together in the representation space by using the IPM should make it easier to generalise the counterfactual outcomes between the two groups.

On the other hand, the goal of the DragonNet is to get rid of all information which is not relevant for confounding due to the outcome and propensity heads in the model. Therefore, the expectation is that for all the different representations obtained for different levels of α, which determines the strength of the treatment prediction objective, the measured confounding bias will be relatively low. Additionally, this low confounding bias in the representations is then expected to lead to relatively good ATE estimates. While all DragonNet models should lead to low biases, we expect that in these low overlap settings, DragonNet models with low treatment assignment weights α in the objective functions should perform slightly better compared to the other versions. Focusing more on predicting the (mostly) extreme propensity scores in favour of outcome prediction is not expected to be a worthwhile trade-off.

The confounding biases measured in the representations and the biases in the final ATE estimates can be found in Figure 4, with the results for both models separately shown in Appendix C. In the first graph it can be seen that all the models which do not try to balance using an IPM, which are all DragonNet models and CFR for $\alpha = 0$, manage to learn unbiased feature representations, as they can fully focus on either accurate outcome prediction, or a combination of both outcome and treatment prediction in the case of DragonNet. These specific models also seem to lead to the ATE estimates with the lowest amount of bias, which can be seen in the graph on the right. When increasing the strength of the IPM in CFR, the confounding bias within the learned representations also seems to increase accordingly. This is also the case in the final ATE estimates, where most of the bias can be explained by the significant confounding bias introduced through the representations. Some additional bias is also present, which most likely comes from the structural overlap violations.

Surprisingly, the DragonNet models with the highest α weights for the treatment prediction objective appear to also lead to the lowest observed biases in the ATE estimates. So, this seems to indicate that keeping some treatment assignment information inside the learned representation may actually be beneficial in these very low overlap settings, which was not initially expected to be the case.

The poor performance of the balancing objective in CFR is somewhat surprising. It was expected to improve the ATE estimation through better generalisation performance when predicting the counterfactual outcomes by moving the treated and control samples closer together in the learned representations, even though it comes at a cost of factual outcome prediction accuracy due to the additional objective. Appendix D shows the IPM and MSE values for the different models, where it can be seen that achieving lower IPM losses also indeed leads to a loss in prediction accuracy, since the MSE values go up. This trade-off between balancing and loss of outcome prediction information, which leads to confounding bias in this case, therefore, does not seem to be worth it in this scenario.

So, the results from this experiment clearly show that using a balancing objective, such as an IPM, when learning representations can lead to a considerable amount of confounding bias in the learned representations. This additional bias in favour of more balancing does not lead to better ATE estimates, indicating that in the specific settings considered here, confounding bias minimisation seems to be a more suitable objective for learning representations. Additional results for the case with non-linear effect modifications can be found in Appendix E. These results are in line with the previous observations, showing that the introduced confounding bias due to balancing also does not seem to be worth the trade-off in that setting.

6.2.2 Estimation using IPW with Learned Representations

In this experiment, the representations obtained using the different learning objectives in the CFR and DragonNet models are used instead as inputs to an IPW estimator, rather than using outcome models for the effect estimation. The propensity scores are calculated using these representations and are then used as weights for the IPW. This could be interesting, as the IPW estimator tends to become quite unstable when using propensity scores obtained from features with very low overlap due to the division by propensity scores close to 0 or 1 during the weighing. By comparing the use of unbiased representations from the DragonNet, or more balanced representations from CFR, we can see which objective could also lead to more suitable representations for IPW estimators. Since extreme propensity scores can lead to very unstable and biased ATE estimates when using weighting estimators, it is expected that the representations which focus only on outcome prediction should lead to the best ATE estimates in this case.

Figure 5 shows the biases obtained using either the CFR representations (left) or the DragonNet representations (right). First, a large amount of variance can be observed for all of the models, which is to be expected for an IPW estimator in such low overlap conditions. Second, it can be seen that, in this case, most of the biases are negative compared to the positive biases in the previous cases. Additionally, for both models, the biases seem to slowly change direction when using the worst overlap setting (at 2 on the x-axis). The negative bias from using the IPW estimator could potentially be due to the combination of the location of the confounder distributions and the specific outcome function used. Appendix F shows the confounder and observed outcome distributions for the treated and control groups. For the control samples, higher values for the covariates lead to smaller propensity weights and higher outcomes due to the outcome function used in this experiment. This leads to rather large values after the weighing because of the division by weights close to 0. Since we estimate the ATE by subtracting the weighed control outcomes from the treated outcomes, such large control values could lead to the negative biases we see in Figure 5.

Finally, it is also worth noting that in the previous experiment, including more treatment assignment information lead to better ATE estimates for the DragonNet. When using these representations for the IPW estimator, it again seems to be the case that the representations which incorporate more treatment assignment information lead to better ATE estimates. This is surprising, as IPW estimators using propensity scores generally don't perform well in such low overlap settings. A possible explanation could be that the propensity scores obtained from these representations are not as extreme as for the original features, since the learned representations still lose some treatment assignment information due to the additional outcome prediction objective.

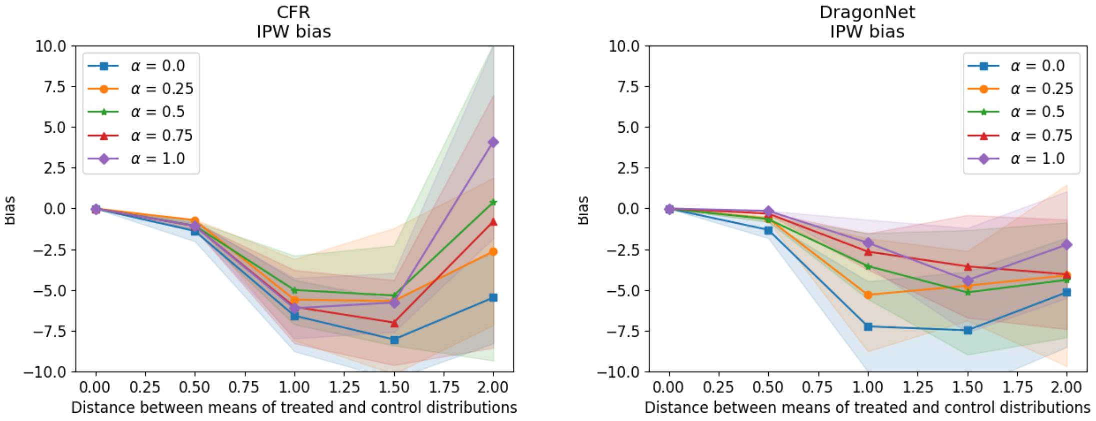


Figure 5: Graphs showing the biases when using the representations obtained by the CFR (left) and DragonNet (right) models as inputs for an IPW estimator. The different lines show the values of α for each model, which correspond to the weights of the IPM or treatment prediction terms in the objective functions for CFR and DragonNet, respectively. Large amounts of variance can be observed, which is to be expected for an IPW estimator under such poor overlap. Additionally, in the case of DragonNet, representations which focus on keeping treatment prediction information seems to perform relatively well compared to the other models.

6.3 Effect of Additional Prognostic and Instrumental Variables

For these experiments, either an additional prognostic or instrumental variable is added to the original non-linear setting used in the previous experiments. For the DragonNet, not much is expected to change compared to the original setting without the extra features. This is due to the outcome and propensity heads which try to get rid of information not related to confounding. In the case of very strong or a larger number of instruments, a high α weight for the treatment prediction objective could potentially lead to a loss of performance, as the model might focus too much on these undesired features. For the CFR, the addition of an instrumental variable could also lead to even worse results, as a strong instrumental variable with very low overlap could be a significant factor in the balancing using the IPM, while the variable actually has no effect on the outcome.

The results for the experiments with the prognostic variable can be found in Appendix G and for the instrumental variable in Appendix H. In both cases, the results are very similar to the ones obtained in the original non-linear setting without the additional variables. This is mainly surprising for the CFR, as the instrumental variables were expected to play a larger role. One interesting thing to note, is that the addition of the instrumental variable leads to the DragonNet models that focus more on treatment prediction to now be worse than the other DragonNet versions, which focus less on treatment prediction, both in terms of bias and variance in the ATE results, which was not the case before. While this was initially expected, the differences are relatively small. For this specific scenario, the lack of significant differences could be due to the effect of one additional instrumental variable not being strong enough compared to the other confounders to show any significant differences in the results.

6.4 Estimating the Covariances from Observed Data

The experiments until this point made use of the potential outcomes and treatments in order to estimate the covariances. The potential outcomes, however, are not observable in practice and can only be used in synthetic data experiments where the true outcome models are known. Therefore, we also experiment with predicting the outcomes under treatment and no treatment, and use these instead of the potential outcomes. Additionally, propensity score models are fitted to obtain treatment predictions to be used in the covariance-based metric. For this experiment, the same linear setting with a hidden confounder from before is used.

Figure 6 visualises the results from running the experiment, which show the confounding biases measured using either the potential or predicted outcomes for increasing strengths of the hidden confounder. The results indicate that it is possible to obtain good estimates of the covariances using only observable data, which means that the covariance-based metric could also potentially be used in real datasets. However, in this setting we considered a well-specified model for the linear outcomes. Model misspecification could lead to worse estimates for the confounding bias, depending on the severity of the misspecification.

7 Discussion

The results from the first experiments show that the bias formulation from D'Amour and Franks (2021) can be used to accurately measure the confounding bias introduced by using some feature representation. However, in order to measure the bias, the space of the representation must be separated into small enough bins to allow for measuring the covariances at different levels of the continuous space. This requires choosing how large of an area within the space should be considered for binning and how many bins should be used. Additionally, the number of samples required to accurately estimate the confounding bias seems to grow as the true amount of confounding bias in the representation increases, which might additionally be affected by the complexity of the specific setting at hand. These binning and sample size choices can affect how accurate the measured confounding biases are, which in turn may affect the conclusions drawn from experiments. Therefore, in the performed experiments, a large number of bins were used with a relatively large number of samples, such that most bins contain enough samples for accurate covariance estimation. However, it has to noted that it is difficult to judge how accurate the measured confounding biases actually were in the

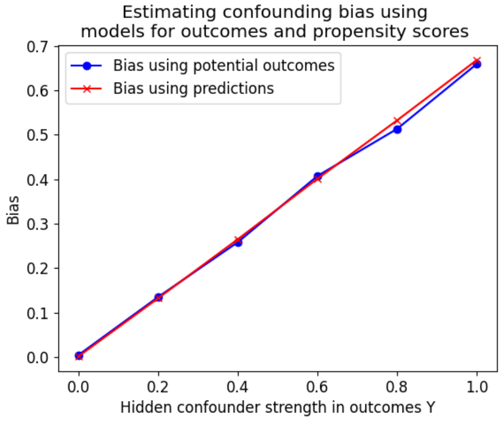


Figure 6: Graph showing the confounding bias when measured using the potential outcomes and treatments (blue), and when using the models to predict the outcomes and propensity scores instead (red). The results are obtained using 10k samples and presented for increasing weights of the hidden confounder on the outcome function, which increases the amount of confounding bias in this setting. The graph shows that the covariance metric can be estimated accurately by using the predicted outcomes and propensity scores.

representation learning experiments due to these issues. But, while there may have been inaccuracies in the estimations, the fact that the exact same estimation settings were used for all models and significant differences in magnitude were observed between the measured biases, it is reasonable to assume that there are indeed significant differences between the different representation methods due to the different representation learning objectives.

The results for the representation learning methods showed clear differences between the balancing objective in CFR and the treatment prediction objective in DragonNet. Across all experiments, the balancing objective introduces a significant amount of confounding bias as the overlap becomes extremely low. The additional balancing that this achieves, however, also does not lead to good ATE estimates. The DragonNet, which manages to consistently show very low levels of confounding bias, also leads to the best ATE estimates in all of the settings. So, the large differences in the obtained biases suggest that confounding bias minimisation could potentially be a more suitable objective for learning representations compared to a balancing objective when estimating average treatment effects in low overlap settings.

From the results for the DragonNet, which manages to consistently achieve low confounding bias in the representations, it appears that including treatment assignment information proves to be beneficial for the ATE estimation. This also seems to be the case when using the representation as input for a weighting estimator. These results are surprising, as D'Amour and Franks (2021) showed in their experiments that using all prognostic information from the confounders, and from other prognostic variables if present, lead to the best weighting estimator results in low overlap settings.

A possible reason for the poor performance of the balancing method could be due to the specific effect of interest here, namely the ATE. CFR with its balancing objective was specifically created by Shalit et al. (2017) to achieve better generalisation capabilities for estimating individual treatment effects through the balancing objective. However, in this work, we were interested in the overall ATE instead of individual treatment effect estimations, which could have lead to the additional generalisation capabilities being less useful. Shi et al. (2019), which introduces the DragonNet, also focuses on ATE estimation and their results showed worse performance by CFR as well. So, those results and the results in this work could indicate that the balancing objective is simply not as suitable for ATE estimation.

While the experiments were performed for a number of different scenarios, such as with various levels of overlap or more complicated, non-linear outcome functions, the results could have changed under certain different settings. For example, if a significantly more complicated outcome function with high dimensional features was used, the increased generalisation performance of CFR could potentially have lead to better results, as the DragonNet could struggle with this in the low overlap scenarios we considered. Furthermore, due to the binning procedure used, we only considered 2-dimensional feature representations, as a higher dimensionality would require significantly more samples to evaluate the conditional covariances. This means that the CFR and DragonNet models had to learn representations with a lower dimensionality than the original features, while they are often used to learn more informative, higher dimensional representations (Shalit et al., 2017; Shi et al., 2019). The CFR model, with its additional balancing objective, may have performed better in such higher dimensional settings for the representations, as the higher capacity could allow for retaining more confounding information, while simultaneously increasing the balance in the representations.

Overall, the covariance-based metric can be useful to understand how different representation learning objectives affect the confounding bias introduced in the learned representations and how this impacts the final ATE estimates. This allows for a better overview of the trade-offs being made in the representations in terms of added bias in favour of some alternative goal, such as increased balance in the representations, which can help with understanding whether such objectives are actually helpful. Although the metric is originally defined in terms of potential outcomes, experiments showed that it can still be estimated from the data, which allows it to potentially be used on real-world datasets to get some understanding of the amount of confounding bias that is introduced.

However, the way of estimating the conditional covariances in this work, by binning the representation space, lowers the applicability of the metric, as only low dimensional representations can feasibly be used without enormous amounts of data. An alternative method for estimating these conditional covariances without binning could significantly increase the use of this metric, allowing it to, for example, also directly be used as an objective in representation learning methods. Such an alternative could be to model the conditional covariances, which would require making parametric assumptions about the variables in the covariances, including the feature representations. This, though, is difficult, or even impossible, to do outside of a very simplistic setting. Exploring non-parametric options for modelling the conditional covariances could be a more suitable approach to increasing the applicability of the covariance metric.

8 Conclusion

The aim of this work was to find a way to measure confounding bias introduced by feature representations and use this to understand the biases introduced by two different representation learning methods for causal inference, namely CFR and DragonNet, in low overlap settings. The covariance-based metric seems to be a suitable method to estimate the amount of confounding bias being introduced by a feature representation. Using this metric, it appears that applying a balancing objective when learning feature representations leads to a large amount of confounding bias, and the additional balance in the representations is not able to lead to better average treatment effect estimates compared to methods that focus on keeping the confounding bias low. So, gaining a better understanding of confounding bias introduced by different methods will allow us to make more confident choices for future causal inferences.

References

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623.

Chernozhukov, V., Newey, W., Quintas-Martınez, V. M., and Syrgkanis, V. (2022). Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR.

Clivio, O., Feller, A., and Holmes, C. C. (2023). Towards representation learning for general weighting problems in causal inference. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

D'Amour, A. and Franks, A. (2021). Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. (2022). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *The Journal of Machine Learning Research*, 23(1):7489–7538.

Melnychuk, V., Frauen, D., and Feuerriegel, S. (2023). Bounds on representation-induced confounding bias for treatment effect estimation. *arXiv preprint arXiv:2311.11321*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

A Proof of Covariance Metric for Confounding Bias

Here, we will provide the proof for the covariance metric as shown in D'Amour and Franks (2021).

First, note that τ^{ATE} can be formulated as

$$\begin{aligned}\tau^{ATE} &= E[Y(1) - Y(0)] \\ &= E[E[Y(1)|d(X)]] - E[E[Y(0)|d(X)]] \\ &= \mu^{(1)} - \mu^{(0)}\end{aligned}$$

Second, we formulate an estimand τ_d^{ATE} for the ATE when conditioning on $d(X)$ as

$$\tau_d^{ATE} = E\left[\frac{TY(1)}{e_d(X)}\right] - \left[\frac{(1-T)Y(0)}{1-e_d(X)}\right]$$

where $e_d(X) = P(T = 1|d(X))$ is the reduced propensity score. Then, we can rewrite both terms:

$$\begin{aligned}E\left[\frac{TY(1)}{e_d(X)}\right] &= E\left[\frac{E[TY(1)|d(X)]}{e_d(X)}\right] \\ &= E\left[\frac{Cov(T, Y(1)|d(X))}{e_d(X)} + \frac{E[T|d(X)]E[Y(1)|d(X)]}{e_d(X)}\right] \\ &= E\left[\frac{Cov(T, Y(1)|d(X))}{e_d(X)}\right] + E[E[Y(1)|d(X)]] \\ &= E\left[\frac{Cov(T, Y(1)|d(X))}{e_d(X)}\right] + \mu^{(1)}\end{aligned}$$

Similar steps can be taken for the other term to find $\left[\frac{(1-T)Y(0)}{1-e_d(X)}\right] = -E\left[\frac{Cov(T,Y(0)|d(X))}{1-e_d(X)}\right] + \mu^0$.

We then obtain $\tau_d^{ATE} = \mu^{(1)} - \mu^{(0)} + E\left[\frac{Cov(Y(1),T|d(X))}{e_d(X)} + \frac{Cov(Y(0),T|d(X))}{1-e_d(X)}\right]$. Taking the difference between τ^{ATE} and τ_d^{ATE} gives the amount of confounding bias introduced by using a feature representation $d(X)$:

$$\tau_d^{ATE} - \tau^{ATE} = E\left[\frac{Cov(Y(1), T|d(X))}{e_d(X)} + \frac{Cov(Y(0), T|d(X))}{1 - e_d(X)}\right]$$

In case $d(X)$ contains all confounding information from the original unconfounded features X, then the unconfoundedness assumption also holds for $d(X)$, i.e., $(Y(0), Y(1)) \perp\!\!\!\perp T \mid d(X)$. This leads to the conditional covariances and, thus, the entire expectation becoming 0, showing that the feature representation does not introduce any confounding bias.

B Confounding Bias With Effect Modification

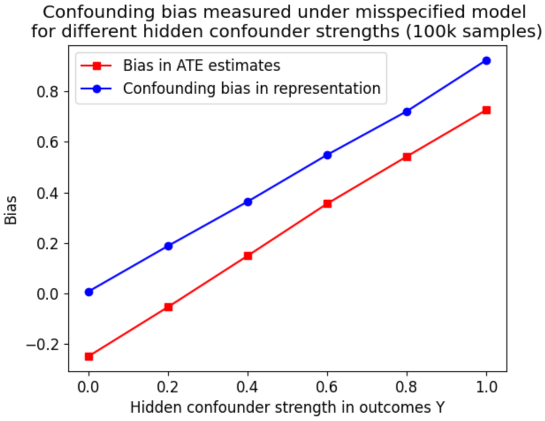


Figure 7: Graphs showing the confounding bias measured in the observed confounders and the bias in the resulting ATE estimate while increasing the weight of the hidden confounder in the outcome model. Additionally, some non-linear effect modification is also present in the outcome model, which causes some model misspecification bias in the outcomes. Here, it can be seen that the ATE bias can be decomposed into confounding bias and model misspecification bias, as for $x = 0$ there is no confounding bias, but a misspecification bias occurs instead in the ATE that persists throughout the graph. Additionally, the combination of these biases may cancel each other out and lead to no overall bias in the ATE (observed around $x = 0.3$).

C Full CFR and DragonNet results in Non-linear Case

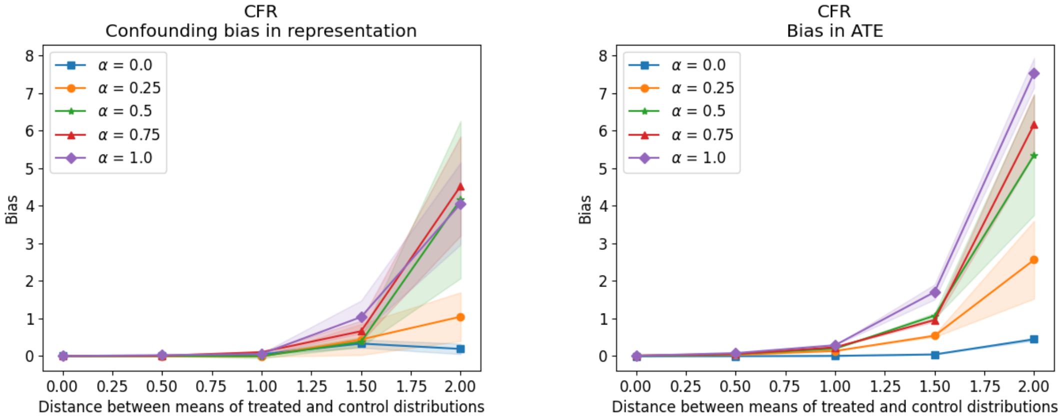


Figure 8: Graphs showing the biases for the CFR models under decreasing levels of overlap. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the IPM terms in the objective functions. The graphs show that more focus on the balancing objective leads to more confounding bias in the representations and in the resulting ATE estimates.

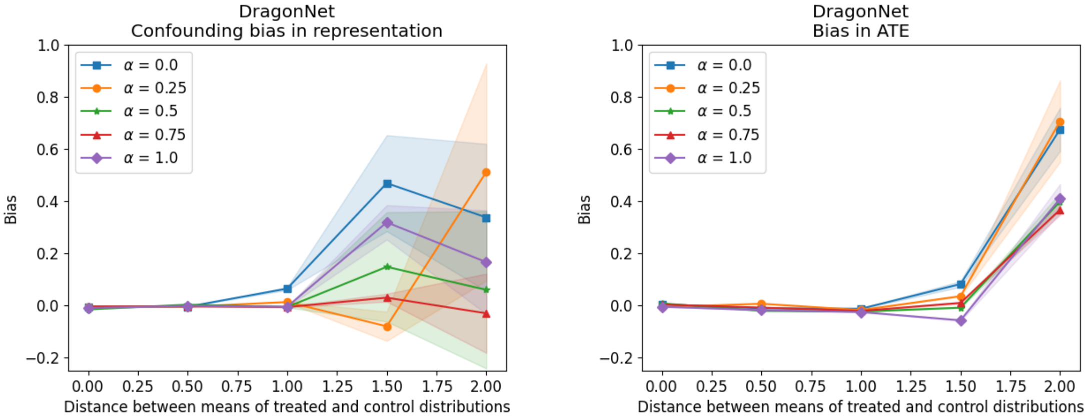


Figure 9: Graphs showing the biases for the DragonNet models under decreasing levels of overlap. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the treatment prediction in the objective functions. Relatively high confounding biases with large variances are measured in the representations, while the ATE estimates don't show these biases. The results could be due to inaccuracies in the covariance estimation.

D CFR Losses in Non-linear Case

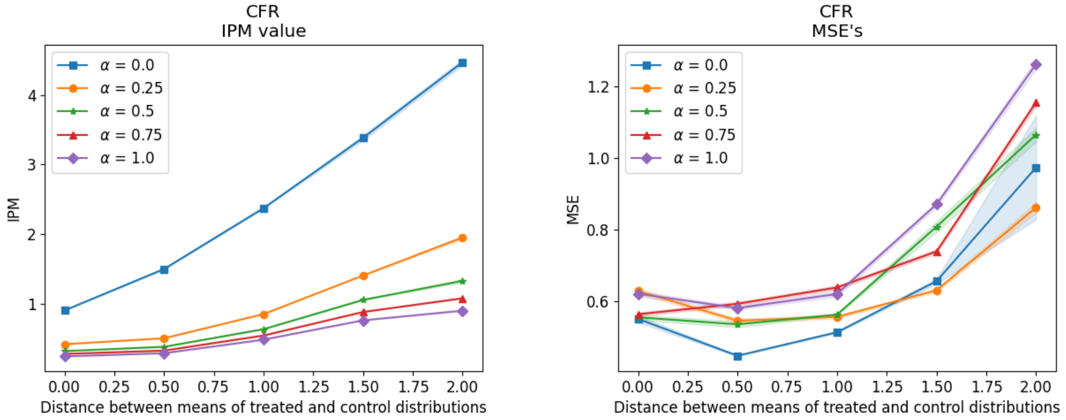


Figure 10: Graphs showing the values of the IPM and MSE's for the CFR models with different IPM strengths α. While a value for the IPM is also shown for the basic TARNet ($\alpha = 0$), the model does not actually use an IPM while training. A trade-off can be seen between lower IPM values, indicating more balance in the representations, and outcome prediction accuracy in the MSE's.

E Full CFR and DragonNet results with Non-linear Effect Modification

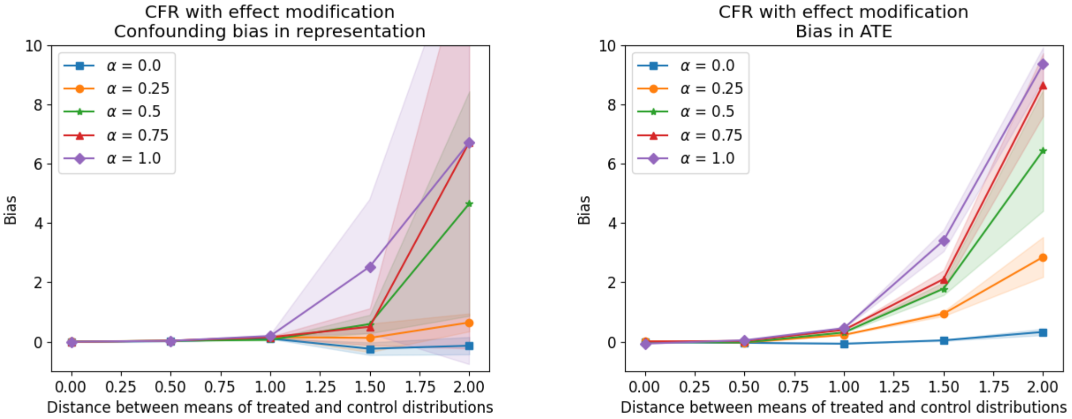


Figure 11: Graphs showing the biases for the CFR models under decreasing levels of overlap with additional non-linear effect modification present. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the IPM terms in the objective functions. The graphs show that more focus on the balancing objective leads to more confounding bias in the representations and in the resulting ATE estimates.

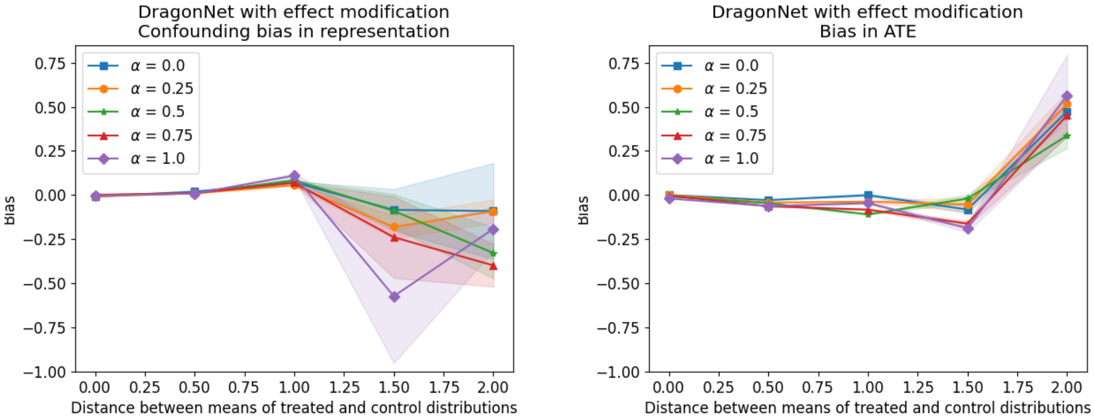


Figure 12: Graphs showing the biases for the DragonNet models under decreasing levels of overlap with additional non-linear effect modification present. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the treatment prediction in the objective functions.

F Example of Covariate and Outcome Distributions in Nonlinear Case

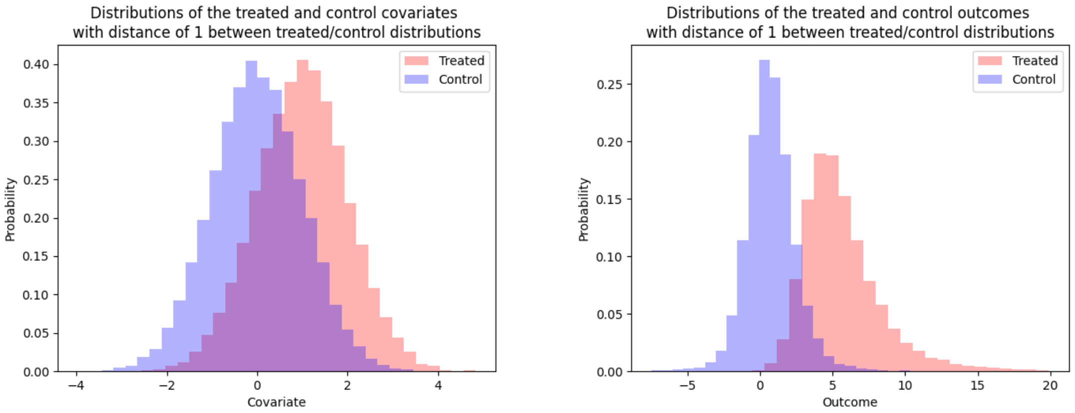


Figure 13: Graphs showing the distributions of the covariates (left) and observed outcomes (right) for the treated and control groups for the IPW experiment. The values are shown for the overlap setting with a distance of 1 between the means of the control and treated distributions for the confounders.

G CFR and DragonNet with Additional Prognostic Variable

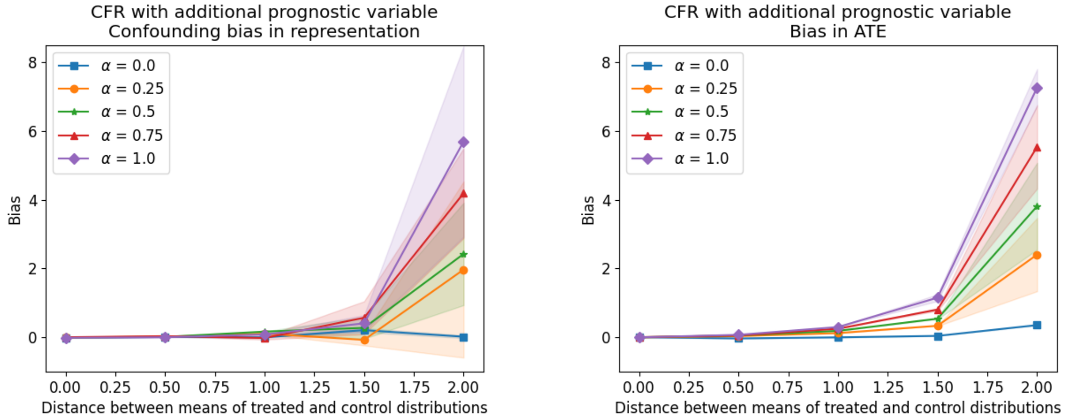


Figure 14: Graphs showing the biases for the CFR models under decreasing levels of overlap with an additional prognostic factor added to features. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the IPM terms in the objective functions.

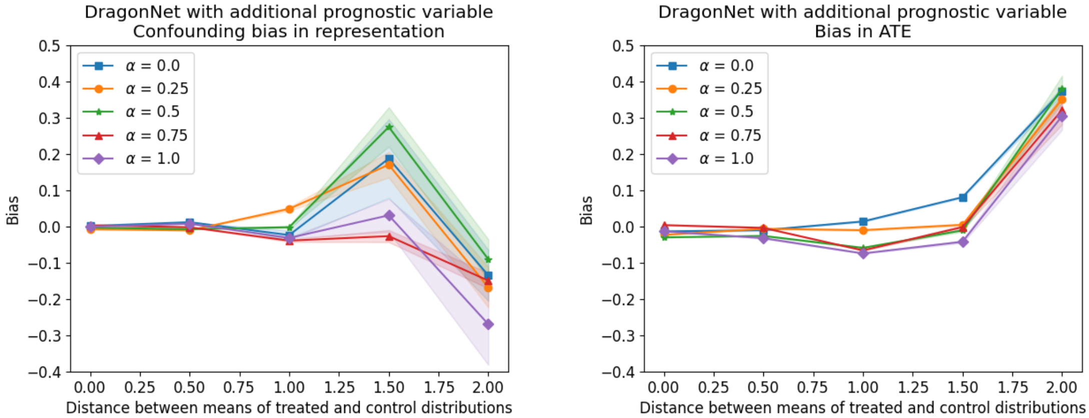


Figure 15: Graphs showing the biases for the DragonNet models under decreasing levels of overlap with an additional prognostic factor added to features. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the treatment prediction in the objective functions.

H CFR and DragonNet with Additional Instrumental Variable

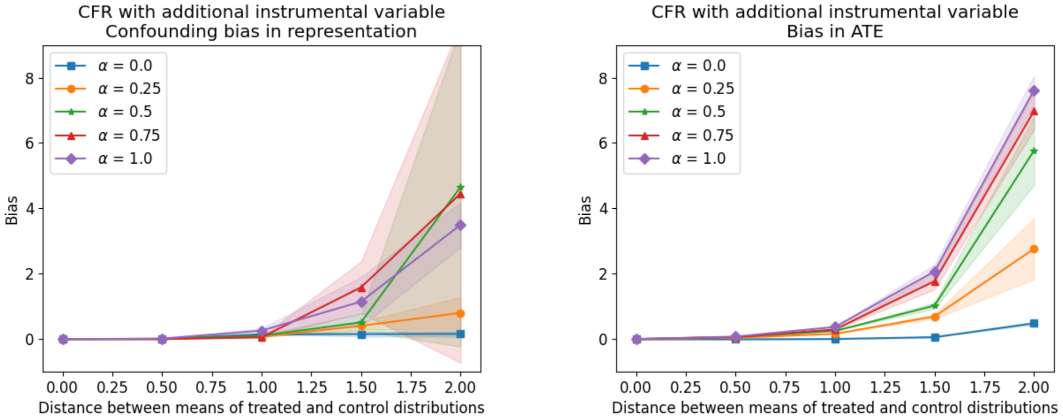


Figure 16: Graphs showing the biases for the CFR models under decreasing levels of overlap with an additional instrumental variable added to features. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the IPM terms in the objective functions.

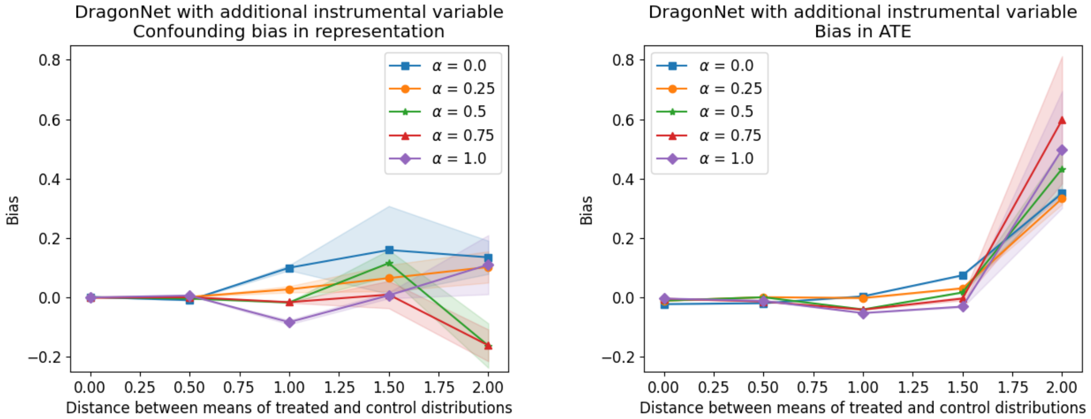


Figure 17: Graphs showing the biases for the DragonNet models under decreasing levels of overlap with an additional instrumental variable added to features. The first graph shows the confounding biases measured within the representations, while the second graphs shows the total biases measured in the ATE estimates. The different lines show the values of α for each model which correspond to the weights of the treatment prediction in the objective functions.