An empirical study of the effects of unconfoundedness on the performance of Propensity Score Matching

Andrej Erdelsky
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

The purpose of this research is to analyze the performance of Propensity Score Matching, a causal inference method for causal effect estimation. More specifically, investigate how Propensity Score Matching reacts to breaking the unconfoundedness assumption, one of its core conceptual pillars. This has been achieved by running PSM on synthetic data that upholds the unconfoundedness condition, and then comparing these results with measurements obtained from running the algorithm on data with confounding features with varying contribution to other variable values and hiding these features individually or in progressively higher numbers. These results are also then compared to Linear Regression, a generic machine learning algorithm, for the sake of comparison of performance. The results obtained point to the observation that when hiding variables that only contribute to the main effect, treatment effect or treatment propensity calculation respectively, PSM performs with the same error no matter which of the three effects the hidden feature affects, making them equivalent in their error contribution. Additionally, it has also become apparent that in all experimental scenarios used in this work, PSM performed very similarly to Linear Regression and did not seem to offer any advantages over the latter in these specific situations.

1 Introduction

The capability to understand causal relations is a difficult computational task essential to many scientific fields. The field of causality has been studied in medical science, economics, epidemiology, and meteorology among others (Guo et al., 2020). The estimation of causal effects has been traditionally done by randomized controlled trials (Cook et al., 2002), but since these are quite often unfeasible in a realistic setting, causal machine learning algorithms for causal effect estimation have become increasingly more popular. Traditional machine learning methods are incapable of detecting these causal relations, but causal algorithms offer a path forward that enables the quantification of the effect that a treatment variable has on an outcome variable, while conditioning on all features of a subject present in the data. To illustrate, let's say the length of an article title affects the click-through rate of said article, the longer the title, the more clicks it gets. But what if the actual reason for the clicks was the quality and renown of certain authors, who coincidentally write longer titles, thus making title length correlated, but not the direct cause of the measured effect on the outcome?

From these examples, it is possible to see that the distinction between actual causation and correlation is crucial. Famously, "correlation doesn't imply causation", but as was discussed, there is also no causation without correlation. The aim of causal effect estimation machine learning algorithms is to specifically address this computational challenge and be able to measure it. Humans can intuitively deduce these relations in day-to-day observations; however, causality is a concept that is hard to define and account for when it comes to machine learning methods because of the complex relationship between correlation and causation.

However, most if not all causal machine learning methods in this field operate ideally only under specific conditions, the main assumptions being "unconfoundedness" and the "overlap assumption". Unconfoundedness of a dataset means that there exist no unmeasured confounders (Guo et al., 2020). In simpler terms, this assumption entails that all features (also known as covariates) that affect treatment and outcome have been observed and measured. The other assumption known as overlap signifies that every subject in the data has a non zero probability of getting either treatment (Rosenbaum and Rubin, 1983).

Because of all these difficulties, this topic can be the subject for complex research, with potential for conflicting viewpoints (King and Nielsen, 2019).

The purpose of this research is to investigate the intricacies of Propensity Score Matching, or "PSM", a causal inference method that allows us to calculate the unbiased estimate of the average treatment effect (ATE) but is often specifically used in the estimation of the average treatment effect for the treated (ATT) (Imbens, 2004). As (Austin, 2011) defines, "propensity score matching entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score". When trying to estimate these causal effects of a specific treatment from data, PSM measures it by comparing a test and control group, that is to say comparing a sample of data-points for which the treatment was "true", with a sample where it was "false". An analogy for this would be to compare the infection rates for a certain virus on patients that got administered a vaccine for it with ones that didn't. On observational data however, there is no guarantee that these two groups are independent of other covariates, implying that the treated and untreated groups often systematically differ in their characteristics (Austin, 2011). In this example, variables like gender, age or genetic predispositions can represent these confounding features, among a multitude of other possibilities.

Propensity Score Matching tries to tackle this issue of group dissimilarity directly by matching data points with the same confounders using propensity scores and then comparing their weighted outcomes. Defined by (Rosenbaum and Rubin, 1983), "the propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates". In other words, the probability of getting treatment is based on observed characteristics. The distribution of measured covariates will be the same between a control and test group with the same propensity scores, allowing the unbiased calculation of the treatment effect through matching. A significant amount of research has been done around this method and multiple implementations of it are also available. A multitude of methods have been tested and used for calculating propensity scores and matching samples, respectively (Lee et al., 2010; Setoguchi et al., 2008; King and Nielsen, 2019).

An important aspect of this topic that will be the focus of this paper is the effect of unconfoundedness on the performance of Propensity Score Matching. As with other causal machine learning methods, unconfoundedness constitutes one of the main key assumptions for PSM to work properly (Rosenbaum and Rubin, 1983) and breaking this assumption, should impact the performance of the algorithm. This work therefore tries to quantify these differences in performance. The methodology will consist of running PSM on synthetic data that upholds the unconfoundedness condition, and then comparing these results with measurements obtained from running the algorithm on data with confounding features with varying contribution to other variable values and hiding these features individually or in progressively higher numbers. These results are also then compared to Linear Regression, a generic machine learning algorithm, for the sake of comparison of performance.

The details of this methodology and related work will be discussed in Section 2 together with a more in-depth explanation of propensity score matching and the specific implementation of it used in this work. Section 3 will discuss the set up and reasoning behind the experiments and the results achieved through them, along with the hypotheses they try to answer. Section 4 will provide further discussion about the implications of the results obtained, while Section 5 will consider the responsibility of the research done. Finally, the research will get its conclusion in Section 6 along with potential paths for further experimentation.

2 Methodology

This section contains the details of the formal setup of the problem setting, along with the explanation of the specific algorithms and models used to achieve experimental results. This is to explain how the approach used helped answer the main question of the paper, namely the impact of unconfoundedness on the performance of Propensity Score Matching as a means of causal effect estimation. Moreover, it also serves as a guide for reproducing the results obtained in later sections.

2.1 Problem Description

All the variables that can be considered when discussing and calculating causal effects are present in figure 1. The effect that every variable type has on the others is critical when calculating the causal effect, which can be viewed as the amount of change that being treated has on a subject, compared to not being treated (Guo et al., 2020). The main effect ($X \rightarrow Y$) and propensity score ($X \rightarrow Z$) add unwanted noise to this value, whereas the causal effect is part of the treatment effect ($Z \rightarrow Y$) along with influences from the features that can be only considered as correlation, not causation. Most importantly, when talking about treatment in this paper, it is always assumed to be binary, meaning each subject either has treatment (Z=1) or doesn't (Z=0).

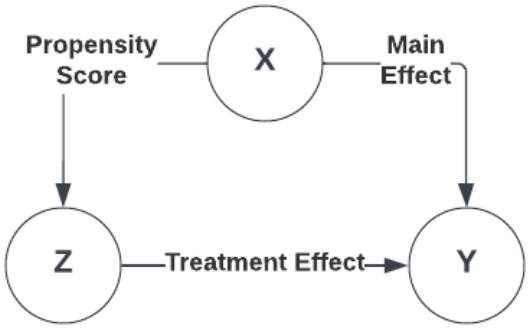


Figure 1: Diagram of Causal Effect, X represents the **features**, Y the **outcome** and Z is the **binary treatment**.

Propensity Score Matching functions by creating matched sets of untreated and treated subjects based on their propensity scores, and then comparing the output Y if they had treatment (Y_1) with the output if they didn't (Y_0) for each of them (Rosenbaum and Rubin, 1983). The Propensity Score : $e_i = \Pr(Z_i = 1|\mathbf{X}_i)$, most often simply referred to as "propensity", is defined as the probability of a subject getting treatment ($Z_i = 1$) based on its set of observed features ($\mathbf{X}_i$) (Rosenbaum and Rubin, 1983). These features are also known as confounding variables or covariates (Guo et al., 2020). Since a single specific subject in the data cannot possibly have an entry with and without treatment, PSM finds an counterfactual subject in a matched group with a similar propensity score, therefore with similar features, and then compares their outcomes.

Just as with other methods relying on the propensity score, for PSM to work, two crucial assumptions need to be upheld. These are unconfoundedness : $(Y(1), Y(0)) \perp\!\!\!\perp Z|X$, and the overlap assumption : $0 < P(Z = 1|X) < 1$ (Austin, 2011). The first assumption means that potential outcomes are independent from the binary treatment assignment conditional on the observed features, this practically means that all features that affect the treatment and outcome have been observed and measured (Austin, 2011). These specific features are often referred to as confounding variables. The latter assumption says that every subject in the data has a non-zero probability of getting treated, meaning that every subject has a potential counterfactual subject in the opposite test group (Austin, 2011). Although both assumptions are important, unconfoundedness is the actual subject of this research.

PSM can accurately output two estimates of causal effect, namely the average treatment effect ATE (1) and the average treatment effect for the treated ATT (2) (Imbens, 2004). Because of time constraints for this research, all experiments analyze ATE because of its ease of use when calculating the ground truth for results and generating synthetic data.

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

$$ATT = E[Y(1) - Y(0)|Z = 1] \tag{2}$$

To answer the main question posed by this paper, the ATE output of PSM when various features are unobserved is compared to its actual true value, which gives an error value. The various experiments conducted in the next section of the paper use different error metrics, these being the Absolute Error (3), the Mean Absolute Error (4), and the Root Mean Squared Error (5).

$$AE = |y_i - x_i| \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - x_i\Big)^2} \tag{5}$$

2.2 Related Work

Following the explanation by (Austin, 2011), there is a multitude of algorithm combinations to consider when utilizing versions of propensity score matching. These can be categorized into the methods used in the acquisition of an accurate propensity score, the numbers in which the pairings found are matched along with the algorithmic way the matching is performed, and finally how the "closeness" of treated and untreated subjects are determined and considered. Each of these components are discussed separately in the ensuing paragraphs.

The most employed technique in the estimation of propensity scores is logistic regression, and it is also the method used in this work. Even though logistic regression is the most frequently used propensity score estimation method seen, bagging, boosting, recursive partitioning, tree-based methods, neural networks, and random forests, among a plethora of others have also been researched for this task (Setoguchi et al., 2008; Lee et al., 2010; McCaffrey et al., 2004).

Next, greedy full matching with replacement is used, a technique discussed at length in (Gu and Rosenbaum, 1993), which signifies that every treatment unit gets matched with n control units and that each control unit gets matched with n treatment units, where n can be chosen (in this case, defaulted to 5). One-to-one matching (1:1) is used the most, but many-to-one matching (M:1) can also be seen. It is also matching with replacement since it is possible to consider a unit more than once when matching them with different units. Finally, the matching is greedy because when choosing specific pairings of units to compare values with when calculating the treatment effect estimation, they are chosen randomly based on their distance of their propensity score instead of optimally. (Gu and Rosenbaum, 1993) has proven that optimal matching does not in fact outperform greedy matching.

To determine this distance and quantify how close units are to each other, the K-nearest neighbors' algorithm has been used. By choosing randomly from a subset of nearest neighbors, we prevent choosing the same unit an abundant number of times when matching, since having discrete values for unit features can cause units to have the same exact propensity score. Moreover, it is also important to mention that no bootstrapping has been used when utilizing this specific version of Propensity score matching.

These decisions about the Propensity Score Matching version specifics used in this paper were motivated by the choice of using the specific code implementation of propensity score matching present in the GitHub repository by (Kelleher, 2018)[1]. Simply put, the choice of this implementation was motivated by its ease of use and the fact it was provided by the supervisors of this research. The minutiae of the implementation of these methods and the choices made can be found in this codebase.

3 Experimental Setup and Results

In this section, the details of every experiment and their setup will be discussed along with the results gathered from them. Each subsection will provide insight on how the results were interpreted and how they address their relevant hypotheses.

A set of three distinct types of experiments has been conducted to try to address all hypotheses from the previous section. These can be distinctly categorized into the **Effect of hiding individual confounding and non-confounding features**, the **Effect of hiding individual features with different effect contributions** and the **Effect of hiding multiple sets of features on synthetic datasets**.

Just as the experiment names indicate, all data used in these experiments is synthetic and therefore generated for the specific purposes of the experiment at hand. The details of this generation will be discussed together with the specific parameters used for each experiment.

3.1 Effect of hiding individual confounding and non-confounding features

The results obtained in this experiment should provide insight into one specific hypothesis, namely that hiding a feature that affects propensity, treatment, and outcome, or in other words, a confounding feature, should impact the performance of PSM. This conversely means that hiding a feature that has no effect on any other variable should theoretically not impact PSM performance.

[1] `https://github.com/akelleh/causality/tree/master/causality/estimation`

3.1.1 Description

This type of experiment consisted of running Propensity Score Matching over multiple iterations on the same common synthetic dataset, each time with different individual features missing. The crucial factor here is that this synthetic dataset has been generated in a way where different sets of covariates are confounders and non-confounding, respectively. Features f_{0-2} are confounders, meaning they affect all effects of the causal graph that can be seen in figure 1, while features f_{3-5} are non-confounding meaning they do not contribute to any other variables but are still present in the data.

By hiding each feature separately, it is possible to observe what happens to the performance of PSM when hiding variables by comparing the obtained results with "baseline" ones that PSM returns when every feature is observed, that is when unconfoundedness holds. Another graph has also been generated that uses Linear Regression instead of PSM. This has been done to be able to compare the reaction to breaking the assumption of unconfoundedness of a causal machine learning algorithm (PSM) with a generic machine learning algorithm (Linear Regression) that hasn't been optimized for causal effect estimation.

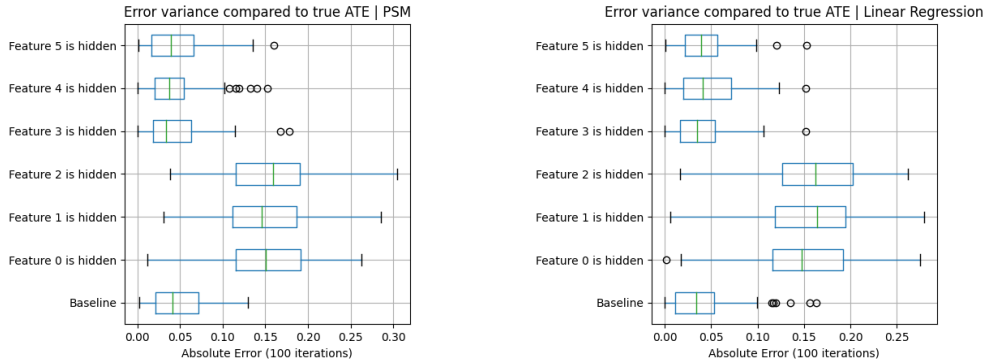
By hiding the appropriate features, it is possible to categorize both graphs into three categories: absolute error when hiding confounding features, absolute error when hiding non-confounding features and finally absolute error when every feature is observed.

The output of PSM that is used here is the ATE, the average treatment effect, which is then compared to the value of the actual causal treatment effect that is utilized when generating the data. The absolute error is then obtained by comparing these two values. Running this over multiple iterations where the dataset is newly generated each time with the same parameters and creating box plots from the results gives a graphical view of the variance in absolute error when hiding specific individual features.

3.1.2 Parameter Setup

Each graph uses the same dataset for calculations and shows results by hiding unique features. Each dataset has a population of 2500, contains 6 features and for each hidden variable test, PSM has been run over 100 iterations on newly generated datasets with the same parameters each time to obtain an accurate absolute error variance as seen on the box-plot graphs. The functions used to generate the dataset are as follows:

- Feature Distribution : $X_i \sim \mathcal{N}(1, 1^2)$
- Main Effect : $x_0 + x_1 + x_2$
- Treatment Effect : $x_0 + x_1 + x_2 + 1$
- Treatment Propensity : $S(x_0 + x_1 + x_2 + \mathcal{N}(0, 1^2))$
- Sigmoid Function : $S(x) = \frac{e^x}{e^x+1}$
- Noise : $\mathcal{N}(0, 1^2)$
- Treatment Function : Binomial distribution $B(1, Propensity)$
- Outcome Function : Main Effect + Treatment Effect * (Treatment - 0.5) + Noise



(a) Using Propensity Score Matching

(b) Using Linear Regression

Figure 2: Error variance compared to true ATE

Again, since there are 6 features present in the data and only features X_{0-2} have been used in the generation of effects, features X_{3-5} are non-confounding. The choice of using a sigmoid function for the treatment propensity calculation, using normal distributions for noise, and using a sum for the feature contributions has been motivated by its use in the GitHub code by (Kelleher, 2018) used in the experiments. Additionally, the feature distribution is centered around 1 in order to not have an expected value of 0 for the feature effects.

3.1.3 Results

On figure 2a, one can see the absolute error variance compared to the true value of ATE when using Propensity Score Matching to estimate the ATE. As is suggested by the data generation function in this specific experiment, Features f_{0-2} are confounded while features f_{3-5} don't have any effect on any other variables. This dichotomy can be clearly seen in this box-plot graph since the amount of error produced by PSM when hiding individual feature is dictated by the fact if that variable is confounded or not.

The baseline error when all features are observed spans between AE values 0 and 0.13 with a mean of 0.04. When confounding features start to become hidden to the algorithm however, the error jumps to AE values spanning from around 0.025 to 0.28 with a mean situated around 0.15, while hiding non-confounding variables doesn't cause any error difference whatsoever compared to the baseline results.

These results therefore confirm the hypothesis that hiding a feature contributing to the main effect, the treatment effect and propensity score calculation impacts the performance of PSM. More specifically, hiding such features causes an average percentage increase in AE of around 275% in this case, while hiding non-confounding features doesn't influence the performance in any noteworthy manner. This is to be expected, since hiding a feature that is confounding effectively prevents PSM to recognize that its effect is only a correlation. This in turn means that PSM interprets the feature's effect as causal, making the estimation wrong by the amount that the hidden feature contributed.

On figure 2b, it is possible to observe that the results acquired using Linear Regression are remarkably similar to the ones obtained by using PSM. These findings can be interpreted

as follows: the PSM offers no discernible advantage compared to generic machine learning algorithms, like Linear Regression, when using it on data where the effect of all features on other variables is homogeneous and a sum of those feature values. Here homogeneous signifies that every feature that is confounded influences the main effect, treatment effect and propensity in the same way, not only specific effects.

3.2 Effect of hiding individual features with different effect contributions

These results should provide insight into three different hypotheses, namely that hiding a feature that only affects the main effect should not impact the performance of PSM, that hiding a feature that only affects the treatment effect should impact the performance of PSM and finally that hiding a feature that only affects the treatment propensity should impact the performance of PSM.

3.2.1 Description

This experiment consists of running Propensity Score Matching on a type of synthetic dataset, while hiding each feature individually over multiple iterations to see how the error changes depending on what feature it is. In this specific dataset, each feature differs in how it contributes to different category of effect in figure 1 (main effect, treatment effect and propensity score). More specifically in figure 3, features f_{0-2} are confounding, while f_3 contributes solely to the main effect, f_4 affects only to the treatment effect and f_5 contributes to the treatment propensity. When hiding each of these variables separately, it should be possible to obtain graphs that show the impact on the performance of PSM when hiding a feature that only affects one specific effect and compare it to the error obtained when hiding a feature that affects all of them.

The error metric used here is the Mean Absolute Error, or MAE, since the ATE output for PSM is compared to its true value when hiding each feature separately. Running this over multiple iterations where the dataset is newly generated each time with the same parameters and creating bar plots from the results should output an accurate graphical view of that error. Just like the previous experiment, another graph has also been generated using Linear Regression instead of PSM for the sake of comparison.

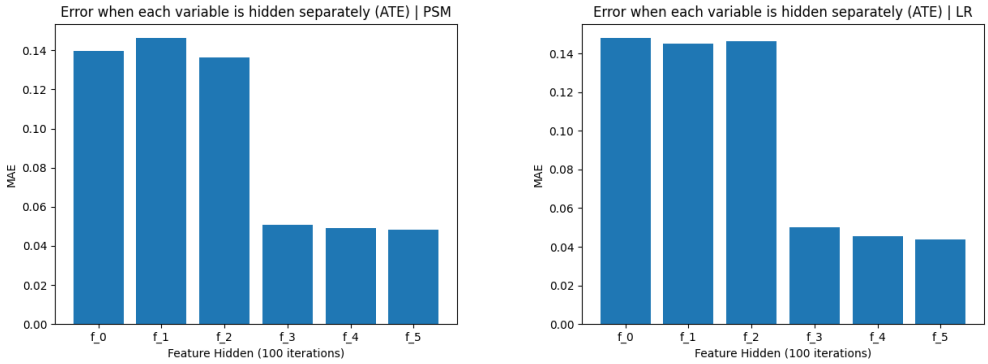
3.2.2 Parameter Setup

The characteristic parameters of the dataset type used is a population of 2500 and the presence of 6 covariant features. However, each feature contributes to effects differently. This can be seen in the functions used to generate the datasets (the other functions are identical to the ones used in section 3.1.2):

- Main Effect : $x_0 + x_1 + x_2 + x_3$
- Treatment Effect : $x_0 + x_1 + x_2 + x_4 + 1$
- Treatment Propensity : $S(x_0 + x_1 + x_2 + x_5 + \mathcal{N}(0, 1^2))$

Moreover, to obtain the results shown in figure 3 experiment type, each bar represents the Mean Absolute Error over 100 newly generated datasets using the same generation parameters to obtain a more accurate representation of an error estimate, instead of specific

anomalies potentially present in unique datasets. The reasoning for the functions used remains the same as the previous subsection experiments (Kelleher, 2018).



(a) Using Propensity Score Matching

(b) Using Linear Regression

Figure 3: Error variance compared to true ATE

3.2.3 Results

On figure 3a, one can see the mean absolute error in ATE when each variable is hidden separately ATE when using Propensity Score Matching. As predicted, removing any of the confounding features f_{0-2} individually causes the same error, namely an MAE of around 0.14. Removing f_{3-5} however results in error values that are significantly lower, more specifically an MAE of around 0.05, which is nearly identical to the mean error value when unconfoundedness holds in the results of section 3.1.3.

These results therefore confirm the hypothesis that hiding a feature contributing only to the main effect should not impact the performance of PSM. Interestingly, the results also disprove the hypotheses that hiding a feature contributing only to the treatment effect or treatment propensity should impact the performance of PSM. Hiding any of these three types of features behaves nearly identically and doesn't cause any major drop in performance for PSM.

These findings can be interpreted as follows: for a feature to cause significant error when hiding it, it needs to affect two or more effects from the main effect, treatment effect and propensity score calculation.

On figure 3b, it is possible to observe that the results acquired using Linear Regression are remarkably similar to the ones obtained by using PSM, meaning that PSM does not provide a discernible advantage compared to Linear Regression in this experimental setting.

3.3 Effect of hiding multiple sets of features on synthetic datasets

These results should provide insight into the last hypothesis, namely that the more hidden variables there are, the worse the algorithm performs. The interesting aspect of this hypothesis is in what manner does PSM worsen its performance with an increasing number of hidden features, and what exactly influences this error trend.

3.3.1 Description

This category of experiments aims to quantify and plot the error in performance of Propensity Score Matching when hiding an increasing number of confounding features. This is achieved by going over the power-set of all feature combinations, grouping them based on size and averaging across the error in each size category. Each line in graphs of figure 4 differs in what dataset was used when calculating the ATE using PSM, and each of these datasets was generated with a different feature function that determines how the features influence the rest of the effects present in figure 1. By having several types of generated datasets, it is possible to obtain graphs that show the impact of hiding an increasing number of features.

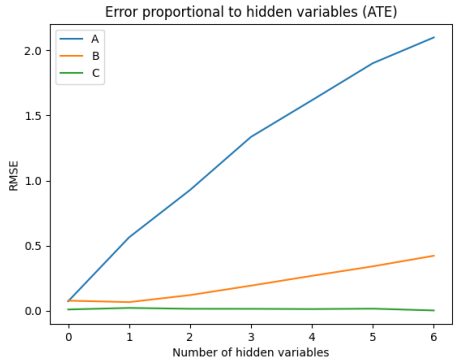
The error metric used in these experiments is the root mean squared error, or RMSE, because the estimated ATE is compared to its true value and averaged over every iteration depending on the size of the subset of features currently being inputted into PSM. This outputs a plot that graphically demonstrates the error trend proportional to the number of hidden variables. Just like the previous experiments, another graph has also been generated using Linear Regression instead of PSM on the same datasets for the sake of comparison.

3.3.2 Parameter Setup

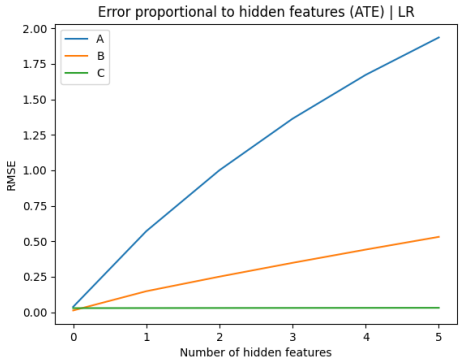
In this series of tests, each line color represents PSM being run on a different dataset. These are distinguished by the specific implementation of the way that all features are utilized in the effects and propensity calculation. All of them, however, have a population of 2500 and contain 6 covariant features. These can be demonstrated by the following generation functions (the other functions are identical to the ones used in section 3.1.2):

- Feature Function : $FF_A : \sum(X_{0-5}) | FF_B : \sum(X_{0-2}) | FF_C : \prod(X_{0-5})$
- Main Effect : $FF(X)$
- Treatment Effect : $FF(X) + 1$
- Treatment Propensity : $S(FF(X + \mathcal{N}(0, 1^2))$

Furthermore, the results shown in figure 4 use RMSE, the root mean squared error, and compare ATE values. The reasoning for the functions used remains the same as the previous subsection experiments (Kelleher, 2018).



(a) Using Propensity Score Matching



(b) Using Linear Regression

Figure 4: Error proportional to hidden features (ATE); A: the feature function used is the sum of all of the features; B: the feature function used is the sum of half of the features; C: the feature function used is the product of all the features.

3.3.3 Results

On figure 4a, one can see the root mean square error in ATE proportional to the number of hidden variables Propensity Score Matching. For 4aA, the feature function used is the sum of all the features. This results in an RMSE value increasing linearly when hiding a progressively larger number of features, starting around 0.5 at 0 features missing and finishing at around 2.1 when all of them are missing. When it comes to 4aB, the feature function used is the sum of half of the features, and the plot follows a similar trajectory than that of 4aA, starting at around 0.1 but ending at around 0.5 when all features are missing. Finally, 4aC uses the product of all the features as a feature function. Here the error stays the same no matter how many of the features are missing, being non-existent.

If the feature values are simply summed and added to the three effects, the error is linearly proportional to the number of unobserved features. When the features are multiplied together and then added to the three features, the error is not dependent on the number of hidden variables since the effect of all features gets amortized into a single value that PSM can easily circumvent when estimating the ATE.

These results therefore confirm the hypothesis that the more hidden variables there are, the worse PSM performs. Removing a progressively larger number of variables and plotting the error forms a straight line with a non-negative slope, meaning the error always increases proportionally to the number of hidden variables. The severity of the slope is dependent on how all the features influence all other variables and how many confounding features there are.

On figure 4b, it is also possible to observe that the results acquired using Linear Regression are once again nearly identical to the ones obtained by using PSM, making this experimental scenario also not indicative of the strengths of Propensity Score Matching.

4 Discussion

To answer the main question of this work, multiple different hypotheses were brought up during experimentation. From the results obtained, it is possible to gain insight into how breaking the unconfoundedness assumption influences the performance of Propensity Score Matching as well as how it fares compared to a non-causal method in the same scenarios.

4.1 Expected Results

Just as expected, when running PSM with missing features, the correctness of the output is clearly dependent on how that feature affects all other variables present in figure 1 as well how many of these features are missing. If individual features are confounders and affect every other variable in the data using the same distribution, removing them individually will result in an output with a significant error compared to the true value.

This is most likely because PSM will not recognize the effect of these features as only correlation since it is not aware of them in the data, and therefore will add the value of this effect to the final causal effect value. The amount by which this faulty estimation is wrong is dependent on how the feature contributions are distributed, but in any case, this error will most certainly cause this value to no longer be accurate.

However, if the feature hidden is non-confounding, the output of PSM will remain the same. One could think that hiding features that do not contribute to anything from the calculation but are still considered by PSM could simplify the estimation and in turn ameliorate the performance by some noteworthy amount, but from experiments done, it is impossible to see any proof of this and it is therefore not the case.

From the last experiment, it is also possible to deduce that removing an increasing number of hidden variables increases the error proportionally to the number of variables hidden in a non-decreasing fashion. The way this error increases and its maximal value is dependent on how many features influence the other effects, by how much they influence them and finally in what way they influence them (i.e. sum, weighted sum, product, etc.). In this sense, PSM behaved as expected.

4.2 Unexpected Discoveries

Most interestingly, two noteworthy findings that contradicted expectations have also been made in this work.

Firstly, when hiding variables that only contribute to the main effect, treatment effect or treatment propensity respectively, PSM performs with the same error no matter which of the three effects the hidden feature affects. One could assume by understanding "the back-door criterion" (Pearl, 2009) that hiding features that affect the treatment effect or propensity calculation should cause a bigger error than hiding a feature solely affecting the main effect. This criterion is the main reasoning behind the unconfoundedness assumption and states the importance of conditioning on all the observed features in the data.

This is done to get rid of the unwanted correlation effect values of these features in the final causal effect estimation, hence the gravity of not having any features hidden. The treatment effect along with propensity directly influences the conditioning of the features and depends on the treatment as opposed to the main effect, which makes these two effects seem more impactful. Based on the results obtained however, there is no difference in the impact they have on PSM performance between them when they are hidden.

Secondly, in the specific experimental scenarios Propensity Score Matching has been utilized, it didn't offer any advantages to a non-causal machine learning algorithm like Linear Regression. In every experiment, Linear Regression performed just as well as PSM and had strikingly equivalent results in every situation when unconfoundedness was broken. This can be attributed to two circumstances.

On one hand, it is possible that the specific experimental scenarios used along with the characteristics of each set of synthetic data created together an environment where the advantages of PSM over traditional machine learning could not be demonstrated. On the other hand, it is also probable that the specific implementation of PSM used in this work is similar enough to Linear Regression to see any notable difference since it uses Logistic Regression to create propensity scores to match each subject. Additionally, the synthetic datasets used in the experiments use linear functions for feature contribution therefore the actual answer is a combination of these two circumstances.

5 Responsible Research

To ensure ethical research and reproducibility of results, several measures have been taken in this work. The first of these measures is that every source of information used, may it be scientific literature or GitHub repositories, has been referenced and given credit to appropriately. Most importantly, every technical aspect of the research needed to reproduce results has also been explained thoroughly.

In the methodology section 2, these technical aspects include baseline mathematical formulas in section 2.1 that are used in later sections of the work, as well as the specific version specifics of the Propensity Score Matching method used in section 2.2. In section 3, every experiment realized in this paper has a description that explains the reasoning behind it, what question the experiment is trying to answer and how it is technically created. In addition to this, all experiments have a parameter setup section that contains the detailing of every variable value and function used, including data generation, while also explaining the motivation behind using them.

Through these means, the transparency of every decision made, and every method used is assured and no information is hidden from the reader. In addition to this, all of the code used in this work to obtain results can be found on the following GitHub repository: (Erdelsky, 2022)[2].

6 Conclusions and Future Work

The purpose of this work was to study the impact of the unconfoundedness assumption on the performance of Propensity Score Matching when estimating causal effects. This was achieved through breaking the assumption in a multitude of ways, by running PSM on synthetic datasets and hiding covariate features in different numbers with varying effect contributions to other variables in the data. In addition to this, the output of Propensity Score Matching has also been compared to Linear Regression, a general machine learning algorithm, in the same experimental scenarios to have a point of reference when interpreting results.

Some of the results acquired met expectations and confirmed preconceived hypotheses, while others unearthed new unforeseen findings. Just as expected, when running PSM with

[2] `https://github.com/Erdandrej/causalityPSM/tree/master/scripts`

missing features, the correctness of the output is dependent on how that feature affects all other variables present as well as how many of these features are missing. More specifically, the more variables the feature effect value affects and the bigger this value is, the more error PSM outputs. This error is also proportional to the amount of such unobserved features and always increases. However, if the feature hidden does not influence any other variable, the output of PSM will remain the same as when every feature is observed by the method.

Interestingly, contrary to prior hypotheses thought of before experimentation, results point to the observation that when hiding variables that only contribute to the main effect, treatment effect or treatment propensity respectively, PSM performs with the same error no matter which of the three effects the hidden feature affects. This information can be therefore interpreted as that all three effects have the same contribution weight to the error when hiding confounding variables. Additionally, it has also become apparent that in all experimental scenarios used in this work, PSM performed very similarly to Linear Regression and didn't seem to offer any advantages over the latter in these specific situations.

From the research and experimentation conducted in this paper, there still exist many avenues of interest to be potentially investigated in the future. Firstly, only the ATE, or the average treatment effect, has been analyzed as output for PSM because of time constraints regarding this work. The ATT, or average treatment effect for the treated, is a metric that is often used when utilizing PSM and it would be worthwhile to investigate the differences in results between it and ATE when reproducing the same experiments. Secondly, it would be beneficial to explore different experimental scenarios when PSM has a clear advantage over Linear Regression since the ones conducted in this paper show these two methods as equivalent. The difference in these scenarios could range from using different versions of PSM to the one used in this work, to employing different functions and distributions with different values for data generation. Lastly, this research would benefit from using real-world causal inference datasets in all its experiments. This would provide results that are more general and realistic, while being independent of artefacts and bias that could arise in a setting where synthetic data is used.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Erdelsky, A. (2022). causalitypsm. `https://github.com/Erdandrej/causalityPSM/tree/master/scripts`.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Kelleher, A. (2018). causality.estimation. `https://github.com/akelleh/causality/tree/master/causality/estimation`.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Pearl, J. (2009). *Causality*. Cambridge university press.

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.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.