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 8, 2022

Abstract

1 Introduction

The capability to understand causal relations is a difficult computational task capable of helping 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 these 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 other features present in the data. To illustrate, how much does getting a vaccine improve the defenses of the immune system against a specific disease, in other words, how much improvement is there in the probability of not getting it? 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.

*A.Erdelsky@student.tudelft.nl
†{S.R.Bongers, J.H.Krijthe}@tudelft.nl

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 variables (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 machine learning algorithm 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 variables, 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) are just some examples of academic papers concerning this topic.

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 the algorithm on data that upholds the unconfoundedness condition, and then comparing these results with measurements obtained from running the algorithm on data with a progressively increasing number of hidden covariates with varying levels of effect contribution to the treatment assignment and outcome, respectively. Moreover, comparing its performance to other, newer methods and extensions that try to achieve similar goals will also provide potentially useful insight.

The details of this methodology 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 is reserved for the contribution that this specific research brings to this scientific field. Section 4 will discuss the set up and reasoning behind the experiments and the results achieved through them, along with their comparison to the previously mentioned hypotheses. Responsible research will then be considered in Section 5, while Section 6 will provide further discussion about the real-world implications of the results obtained. Finally, the research will get its conclusion in Section 7 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 ($X \to Y$) and propensity score ($X \to Z$) add unwanted noise to this value, whereas the causal effect is part of the treatment effect ($Z \to Y$) 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).

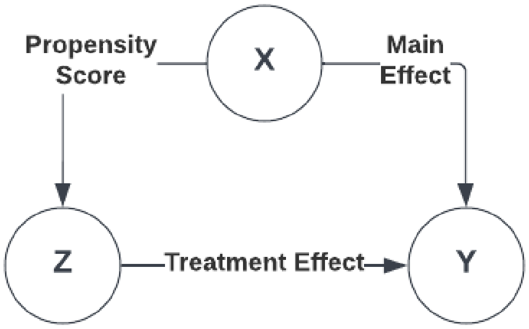


Figure 1: Diagram of Causal Effect, X represents the **features**, Y the **outcome** and Z is the **binary treatment**.

As already stated in the Introduction, this work investigates the impact of unconfoundedness on Propensity Score Matching, in other words, how much error does PSM have when features are hidden to the algorithm. 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 (Y1) with the output if they didn't (Y0) for each of them (Rosenbaum and Rubin, 1983). The Propensity Score (1), most often simply referred to as "propensity", is defined as the probability of a subject getting treatment based on its observed features (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.

$$e_i = Pr(Z_i = 1|\mathbf{X}_i) \tag{1}$$

Just as with other methods relying on the propensity score, for PSM to work, two crucial assumptions need to be upheld. These are unconfoundedness (2) and the overlap assumption (3) (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.

$$(Y(1), Y(0)) \perp\!\!\!\perp Z|X \tag{2}$$

$$0 < P(Z = 1|X) < 1 \tag{3}$$

PSM can accurately output two estimates of causal effect, namely the average treatment effect ATE (4), the average treatment effect for the treated ATT (5) (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.

$$ATE = E[Y(1) - Y(0)] \tag{4}$$

$$ATT = E[Y(1) - Y(0)|Z = 1] \tag{5}$$

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 (6), the Mean Absolute Error (7), and the Root Mean Squared Error (8).

$$AE = |y_i - x_i| \tag{6}$$

$$MAE = (\frac{1}{n})\sum_{i=1}^{n} |y_i - x_i| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - x_i\Big)^2} \tag{8}$$

2.2 Hypotheses

To understand the reasoning behind the experiments conducted, it is important to state the hypotheses that were formulated prior to experimenting. These aimed to answer the research question concerning the effects of unconfoundedness and were thought of while researching the specific subject before engaging with any experimentation. Moreover, addressing each hypothesis entails providing a satisfactory answer for the main question of the paper. The considered hypotheses include:

- A hidden feature that **only** affects **the main effect should not** impact the performance of PSM.
- A hidden feature that **only** affects **the treatment effect should** impact the performance of PSM.
- A hidden feature that **only** affects **the propensity score should** impact the performance of PSM.
- A hidden feature that affects **propensity, treatment and outcome should** impact the performance of PSM.
- The **more hidden variables** there are, the **worse the algorithm** performs.

2.3 Related Work

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 (Austin, 2011). Each of these components are discussed separately in the ensuing paragraphs.

The most employed technique in the estimation of propensity scores is logistic regression (Austin, 2011), 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; Austin, 2011).

Next, greedy full matching with replacement is used, a technique discussed at length in (Gu and Rosenbaum, 1993), which signifies that each unit in the data gets matched with n units from the opposite group, 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 (Austin, 2011). 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]. 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 the 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 different feature combinations with a unique synthetic dataset**, the **Effect of hiding different individual features with different synthetic datasets** 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 different feature combinations with a unique synthetic dataset

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 dataset is considered "unique" since all graphs present in figure 2 have been constructed by running PSM on this single dataset. The crucial factor here is that this synthetic dataset has been generated in a way where different sets of covariates are confounded and non-confounding, respectively. Features X_{0-2} are confounded, meaning they affect all effects of the causal graph that can be seen in figure 1, while features X_{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 a generic Linear Regression machine learning algorithm 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.

[1] `https://github.com/akelleh/causality/tree/master/causality/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 results obtained in this experiment should provide insight into one specific hypothesis, namely that hiding a feature that affects propensity, treatment and outcome 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.

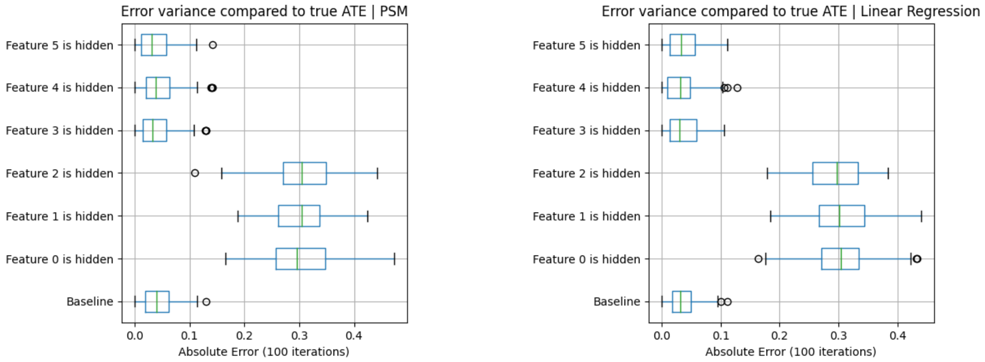
The output of PSM that is used here is the ATE, the average treatment effect, which is then compared to the value of the actual causal treatment effect that is utilized when generating the data. The absolute error is then obtained by comparing these two values. Running this over multiple iterations where 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 different 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 : $X \sim 0.5 * \mathcal{N}(0, 1^2)$
- Main Effect : $x_0 + x_1 + x_2$
- Treatment Effect : $x_0 + x_1 + x_2 + 1$
- Treatment Propensity : $Sigmoid(x_0 + x_1 + x_2 + \mathcal{N}(0, 1^2))$
- Sigmoid Function : $Sigmoid(x) = \frac{e^x}{e^x+1}$
- Noise : $\mathcal{N}(0, 1^2)$
- Treatment Function : $\binom{1}{Propensity}$
- Outcome Function : Main Effect + Treatment Effect * Treatment + Noise

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 features and 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. Moreover, the constant of 1 present in the treatment effect equation represents the true value of the causal effect, meaning the correct ATE desired is in fact 1. This value is then compared to the ATE value returned by PSM and Linear Regression, which results in the absolute error seen in the box plots on figures 2.



(a) Using Propensity Score Matching

(b) Using Linear Regression

Figure 2: Error variance compared to true ATE

3.1.3 Results: Propensity Score Matching

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 X_{0-2} are confounded while features X_{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 0.1 with a mean of 0.05. When confounding features start to become hidden to the algorithm however, the error jumps to AE values spanning from around 0.15 to 0.45 with a mean of 0.3, 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 500% in this case, while hiding non-confounding features doesn't influence the performance in any noteworthy manner.

3.1.4 Results: Linear Regression

On figure 2b, one can see the absolute error variance compared to the true value of ATE when using Linear Regression to estimate the ATE. Here, it is possible to observe that the results are remarkably similar to the ones obtained by using PSM, if not identical, albeit with more pronounced and varied outliers in the results (values represented by black circles, outside of 75% of obtained results).

These findings can be interpreted as follows: the causal machine learning algorithm 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 different individual features with different synthetic datasets

3.2.1 Description

These experiments consist of running Propensity Score Matching on many datasets that differ in how their features contribute to different effects, while hiding each feature individually over multiple iterations to see how the error changes depending on what feature it is. The difference between datasets is represented by the last feature, as in every other feature contributes to every category of effect in figure 1 (main effect, treatment effect and propensity score) except one, where only the last feature X_4 contributes. For example, PSM is run on a dataset where every feature but the last one contributes to the main effect, treatment effect and propensity, and where the last feature solely contributes to the treatment effect. When hiding each of these variables separately, it should be possible to observe the impact on the performance of PSM when hiding a feature that only affects treatment effect.

By having several types of generated datasets, it is possible to obtain graphs that show the impact of hiding a variable that solely influences the main effect, treatment effect and propensity score calculation individually. A dataset where every feature contributes to every effect is also used, to serve as a baseline comparison for the other results in the graph. 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.

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.

3.2.2 Parameter Setup

The experiments conducted here differ in the generated dataset that PSM was run on. In figure 3, each bar color (A, B, C and D) corresponds to a different type of dataset. The common characteristic parameters between them is a population of 2500 and the presence of 5 covariant features. However, they are categorized by the way their last feature, here f_4, influences the main effect, treatment effect and propensity score respectively. This can be seen in the functions used to generate the datasets:

- Feature Distribution : $X \sim 0.5 * \mathcal{N}(0, 1^2)$
- Main Effect : $x_0 + x_1 + x_2 + x_3$ ($+\ x_4$ if A or B)
- Treatment Effect : $x_0 + x_1 + x_2 + x_3 + 1$ ($+\ x_4$ if A or C)
- Treatment Propensity : $Sigmoid(x_0 + x_1 + x_2 + x_3$ ($+\ x_4$ if A or D) $+\mathcal{N}(0, 1^2))$
- Sigmoid Function : $Sigmoid(x) = \frac{e^x}{e^x+1}$
- Noise : $\mathcal{N}(0, 1^2)$
- Treatment Function : $\binom{1}{Propensity}$

- Outcome Function : Main Effect + Treatment Effect * Treatment + Noise

Moreover, to obtain the results shown in figure 3 experiment type, each bar color represents the Mean Absolute Error over 100 different 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, along with the true ATE being equal to 1 (Kelleher, 2018).

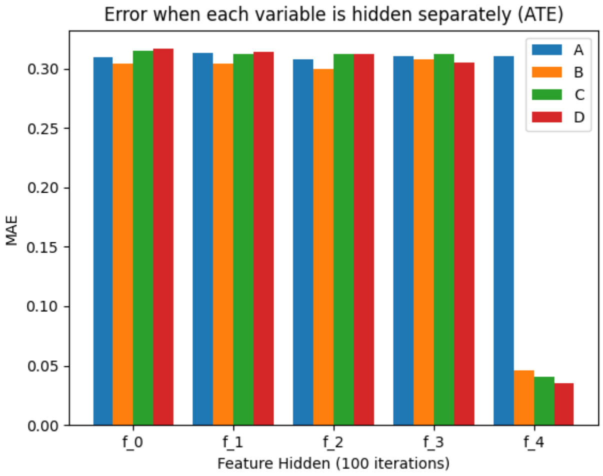


Figure 3: Error when each feature is hidden separately (ATE); A : baseline dataset where every feature contributes to every effect, B : dataset where the last feature contributes solely to the main effect, C : dataset where the last feature contributes solely to the treatment effect, D : dataset where the last feature contributes solely to the treatment propensity calculation

3.2.3 Results A : Baseline

On figure 3A, one can see the mean absolute error in ATE when each variable is hidden separately ATE when using Propensity Score Matching with all features observed. As predicted, removing any feature individually causes the same error, namely a MAE of around 0.3. This knowledge will be used as a comparison for the other results.

3.2.4 Results B : Main Effect

On figure 3B, one can see the mean absolute error in ATE when each variable is hidden separately ATE when using Propensity Score Matching where the last feature f_4 solely contributes to the main effect. Removing f_{0-3} results in similar MAE values as in 3A but removing f_4 causes an error that is significantly lower, namely a MAE of around 0.05.

These results therefore confirm the hypothesis that hiding a feature contributing to the main effect should not impact the performance of PSM. More specifically, hiding such features doesn't cause any major drop in performance for PSM.

3.2.5 Results C : Treatment Effect

On figure 3C, one can see the mean absolute error in ATE when each variable is hidden separately ATE when using Propensity Score Matching where the last feature f_4 solely contributes to the treatment effect. Removing f_{0-3} results in similar MAE values as in 3A but removing f_4 causes an error that is significantly lower, namely a MAE of around 0.045.

These results therefore disprove the hypothesis that hiding a feature contributing to the treatment effect should impact the performance of PSM. Hiding such features doesn't cause any major drop in performance for PSM, it behaves nearly identically to removing a feature that solely contributes to the main effect.

3.2.6 Results D : Treatment Propensity

On figure 3D, one can see the mean absolute error in ATE when each variable is hidden separately ATE when using Propensity Score Matching where the last feature f_4 solely contributes to the treatment propensity. Removing f_{0-3} results in similar MAE values as in 3A but removing f_4 causes an error that is significantly lower, namely a MAE of around 0.04.

These results therefore disprove the hypothesis that hiding a feature contributing to the treatment propensity should impact the performance of PSM. Hiding such features doesn't cause any major drop in performance for PSM, it behaves similarly to removing a feature that solely contributes to the main effect or treatment effect. Interestingly, removing a feature that contributes only the propensity causes the least amount of error, followed by treatment effect (12.5% increase in MAE) and finally main effect (25.0% increase in MAE).

These findings can be interpreted as follows: in order 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.

3.3 Effect of hiding multiple sets of features on synthetic datasets

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 graph 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. 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.

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.

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 5 covariant features. These can be demonstrated by the following generation functions:

- Feature Distribution : $X \sim 0.5 * \mathcal{N}(0, 1^2)$
- 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 : $Sigmoid(FF(X + \mathcal{N}(0, 1^2))$
- Sigmoid Function : $Sigmoid(x) = \frac{e^x}{e^x+1}$
- Noise : $\mathcal{N}(0, 1^2)$
- Treatment Function : $(\frac{1}{Propensity})$
- Outcome Function : Main Effect + Treatment Effect * Treatment + Noise

Furthermore, the results shown in figure 4 uses RMSE, the root mean squared error and compare ATE values. The reasoning for the functions used remains the same as the previous subsection experiments, along with the true ATE being equal to 1 (Kelleher, 2018).

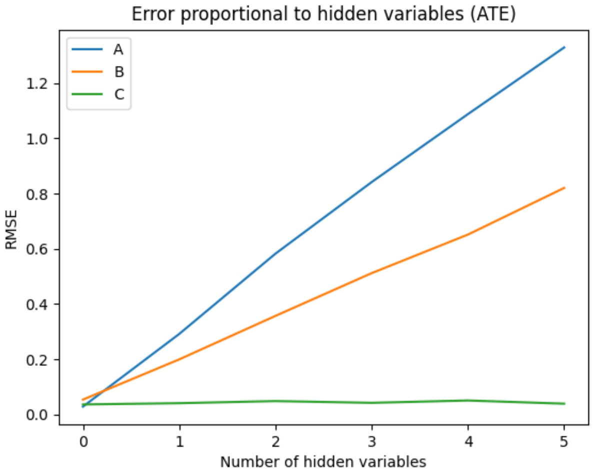


Figure 4: Error proportional to hidden variables (ATE); A : the feature function used is the sum of all of the features; B : the feature function used is the sum of half of the features; C : the feature function used is the product of all the features.

3.3.3 Results

On figure 4, one can see the root mean square error in ATE proportional to the number of hidden variables Propensity Score Matching. For 4A, 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 at 0 features missing and finishing at around 1.2 when all of them are missing. When it comes to 4B, the feature function used is the sum of half of the features, and the plot follows a similar trajectory than that of 4A, starting a little worse but ending at around 0.8 when all features are missing. Finally, 4C uses the product of all the features as a feature function. Here the error stays essentially the same no matter how many of the features are missing.

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 disprove the hypothesis that the more hidden variables there are, the worse PSM performs. The error proportional to hidden variables is clearly dependent on how all the features influence all other variables, meaning the error doesn't necessarily get worse, since it can stay identical independent of how many features are missing.

4 Discussion

5 Responsible Research

6 Conclusions and Future Work

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.

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.

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.