Assignment-Control Plots: A Visualization tool for Causal Inference Studies

Rachael C. Aikens, Michael Baiocchi

1 Abstract

An important step for many causal inference study designs is ensuring that the compared treated and control observations are similar in terms of their baseline measured covariates. However, not all baseline variation is equally important. In the observational context, balancing on baseline variation summarized in a propensity score can help reduce bias due to self-selection. In both obeservational and experimental studies, controlling baseline variation associated with the expected outcomes can help increase the precision of causal effect estimates. In some study designs, it can be essential to correctly characterize and efficiently use an instrumental variable, a treatment effect modifier, or a forcing variable. We propose a set of visualization tools which reduce the space of measured covariates into a set of axes most important to a causal question. These "assignment-control plots" and variations thereof may be a useful diagnostic and educational tool for a wide variety of causal inference study designs.

2 Introduction

A fundamental problem of causal inference is the impossibility of observing counterfactuals: Once an individual has recieved a treatment or exposure, we can never observe what might have happened to that individual had they not recieved treatment, and vice versa. Most studies of causal inference attempt to address this problem by comparing samples of treated individuals with samples of untreated individuals, ideally in some setting which controls for ways that these two groups might systematically differ. Intuitively, one can imagine that our observations of the control sample are used as approximations in order to understand what *might* have happened to the treated individuals had they been untreated. A question fundamental to these approaches is: How would we like our compared treated and control samples to be similar in order to obtain a clear understanding of the causal effect? Said another way, what aspects of baseline variation are most important to address estimate the causal effect with minimal bias and variance? In the randomized experimental setting, we would like our compared groups to be balanced in terms of variation important to the outcome, since this reduces the variance of our estimates. In the observational setting, we have the additional burden of correcting for bias-inducing imbalances stemming from an unknown assignment mechanism.

Researchers have proposed a variety of matching and conditioning methods to address baseline variation between treated and control samples variety of observational contexts. One popular approach is subclassification or adjustment on an estimated propensity score, which summarizes the measured baseline variation influencing the probability of assignment. Intuitively, propensity score methods attempt to model the treatment assignment mechanism based on observed covariates, so that it can be adjusted for. Under suitable assumptions, matching exactly on the propensity score recapitulates a completely randomized controlled experiment, allowing for identification and unbiased estimation of treatment effect. However, conditioning solely on the propensity score may neglect baseline variation which is unimportant for treatment assignment but influential on the *potential outcomes*, potentially resulting unfavorably high variance and low power. For this reason, propensity score matching has been critiqued as statistically inefficient compared to other

methods which optimize for a more comprehensive form of covariate similarity between matched sets (King and Nielsen 2016).

In light of these critiques, methods such as Mahalanobis distance matching - which seek to establish similarity between matched sets over *all* measured covariates - seem appealing. However, not all measured covariates are necessarily important to the causal problem - especially as researchers collect more and more comprehensive observational data with many measured covariates. The less-commmonly known prognostic score, formalized by Hansen (2008), models the expected outcome of each subject under the control assignment, based on the observed covariates. Matching or conditioning on the prognostic score reflects the ideal of controlling for baseline variation influencing the potential outcomes under the control assignment. Interestingly, under suitable assumptions, balancing on the prognostic score results in a form of covariate balance which leads to unbaised estimation of the causal effect, analogous to the propensity score. The primary difference is that the propensity score controls for baseline variation influencing treatment assignment, whereas the prognostic score controls for baseline variation influencing the potential outcome under the control assignment. A small but growing body of literature suggests that methods which match jointly on a prognostic score and a propensity score may be a favorable approach in some observational contexts, optimizing directly for propensity score balance (which reduces bias), and for prognostic score balance (which reduces bias as well as variance and increases power in analyses of gamma sensitivity)(Leacy and Stuart 2014; Antonelli et al. 2018; Aikens, Greaves, and Baiocchi 2020).

Estimating the prognostic score in a way that does not overfit is a somewhat nuanced problem discussed in detail by Aikens, Greaves, and Baiocchi (2020) and summarized in section 3.3. However, once a prognostic model is fit, it is possible to generate assignment control plots (indroduced in Aikens, Greaves, and Baiocchi (2020)), which visualize each subject in an observational dataset according to their propensity and prognostic scores. These are two (often interrelated) features which are directly relevant to observational studies of causality: propensity score similarity between compared individuals reduces bias, while prognostic score similarity between compared individuals reduces bias as well as variance and increases power in sensitivity analyses for unobserved confounding. While assignment-control plots are a relatively minor point in this original study, here we will describe a more detailed investigation of the uses of assignment-control plots in visualization, as well possible extensions to instrumental variable studies and regression discontinuity. We suggest that assigment control plots and variations thereof may be a useful visualization and teaching tool in many branches of causal inference research.

3 Methods

3.1 Notation and Background

We adopt the Neyman-Rubin potential outcomes framework, in which a sample is described by

$$\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$$

where the triplet (X_i, T_i, Y_i) describes an individual with measured covariates X_i, binary treatment assignment indicator T_i, and observed outcome Y_i. We take $Y_i(T)$ to represent the potential outcome of individual i under treatment assignment T. The fundamental problem of causal inference is that it is impossible to observe both potential outcomes, $Y_i(0)$ and $Y_i(1)$ for any individual.

The propensity score is defined as $e(X) = P(T = 1|X)$. The popularity of the propensity score in observational studies stems primarily from its use as a balancing score, i.e.

$$T \perp X | e(X)$$

.

That is, within level-sets of the propensity score, the treatment assignment is independent of the measured covariates. Under the assumption of strongly ignorable treatment assignment, exact matching on the propensity score allows for unbiased estimation of the treatment effect (Rosenbaum and Rubin 1983).

The prognostic score is defined by Hansen as any quantity $\Psi(X)$ such that

$$Y(0) \perp X | \Psi(X)$$

Inessense, a prognostic score is any function of the measured covariates which – through conditioning – induces independence between the potential outcome under the control assignment and the measured covariates. It is thus, by definition, a balancing score as well. Under regularity conditions analogous to those for the propensity score, conditioning on the prognostic score also allows for unbiased estimation of the treatment effect, as described in more detail by Hansen (2008). When $Y(0)|X$ follows a generalized linear model $\Psi(X) = E[Y(0)|X]$. In the literature, the prognostic score is often treated more informally as a model for the expected outcome under the control assignment given the observed covariates.

3.2 Set-up

The results that follow depict several simulated datasets. The primary generative model for these is based on Aikens, Greaves, and Baiocchi (2020) and is specified as follows:

$$\begin{aligned} X_i &\sim Normal(0, I_{10}) \\ T_i &\sim Bernoulli\left(\frac{1}{1 + exp(\phi(X_i))}\right) \\ Y_i(0) &= \tau \Psi(X_i) + \epsilon_i \\ \epsilon_i &\sim N(0, 1) \end{aligned}$$

where $\phi(X)$ and $\Psi(X)$ represent the true propensity and prognostic score functions, given by

$$\begin{aligned} \phi(X_i) &= c_1 X_{i1} - c_0 \\ \Psi(X_i) &= \rho X_{i1} + \sqrt{(1 - \rho^2)} X_{i2}. \end{aligned}$$

Where c_1, c_0, and ρ are constants. In particular, the form for the prognostic function above guarantees that $\rho = Corr(\phi(X), \Psi(X))$.

3.3 Fitting the Score Models

In real observational studies, the propensity and prognostic score models are not known. Conventionally, the propensity scores are often estimated from a logistic regression of the baseline covariates on treatment assignment, fit on the entire study sample. Fitting the prognostic score may be somewhat more nuanced (Aikens, Greaves, and Baiocchi (2020)). First, since the prognostic model is meant to predict the outcome under the control assignment, the prognostic model is fit only on controls. Thus, all prognostic score estimates on the treatment population are necessarily extrapolations. Second, fitting the prognostic model on the entire control population raises concerns of overfitting (Hansen 2008; Abadie, Chingos, and West 2018). In order to avoid concerns of overfitting and preserve the separation of the design and analysis phases of the study, Aikens, Greaves, and Baiocchi (2020), propose a Pilot Design, in which a subset of the control individuals is selected and held aside for the purpose of fitting the prognostic model. These controls - comprising a "Pilot data set" are then discarded, so that the observational units used to train the prognostic model are dijoint

from the set used in the final analysis. The question of how to optimally select the control observations for the pilot set is a difficult one, described in more detail elsewhere (Aikens, Greaves, and Baiocchi 2020; Aikens et al. 2020).

In order to clearly demonstrate the ideas behind the assignment control plots in this paper, most of the examples that follow bypass the problem of score estimation by using the ground-truth propensity and prognostic scores, as specified in our simulations. For more discussion on the realities of fitting the score models, see Aikens, Greaves, and Baiocchi (2020)

4 Results

4.1 Assignment-Control Plots and Data Diagnostics

A first use for the assignment-control plot is as an informal data diagnostic. Researchers running observational studies often want a basic understanding of the baseline variation between the treatment and control groups before they begin. Plotting the standardized mean differences between the treatment groups is a common starting place to understand imbalances between groups, but when there are many covariates – some of which are irrelevant – it can be difficult to tell by eye which imbalances should be of most concern. Histograms of propensity score overlap can be an important diagnostic for checking that the treatment and control groups overlap in terms of their probability of treatment (since this condition is essential for many causal inference methodologies). However, the propensity score does not necessarily reflect all aspects of covariate balance which may be important.

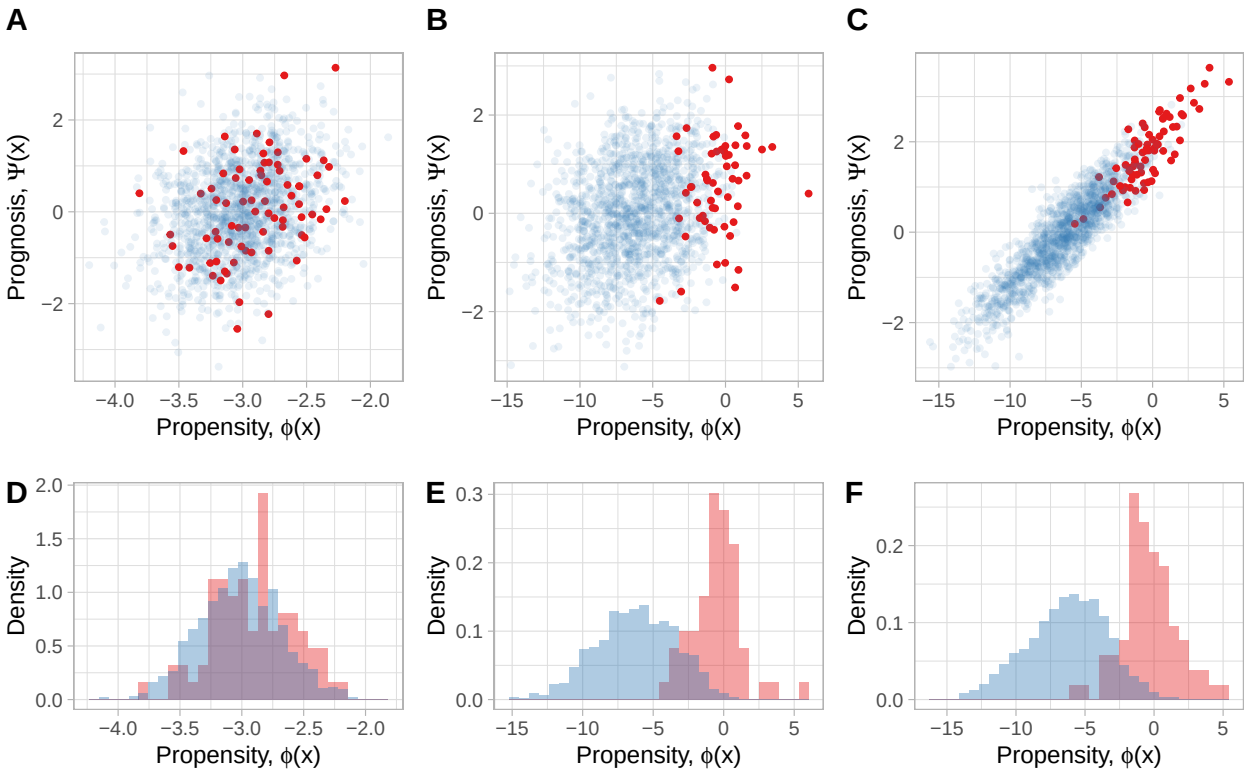

Figure 1: Assignment-Control plots (A-C) and propensity score density histograms (D-F) for three simulated observational datasets. Red points represent treated observations, blue points represent control.

Figure 1 shows example assignment-control plots (A-C) and propensity score density histograms (D-F) for three simulated observational datasets. It is worth noting that the propensity score histogram for any

dataset simply shows the marginal distribution of the data clouds shown in any assignment control plot. Panels A and D in Figure 1 depict the assignment-control plots and propensity histograms for a favorable observational dataset. There is substantial overlap between the treated and control subjects both in terms of their propensity for treatment and their likely outcome under the control assignment. The correllation between treatment and prognosis is low. A researcher viewing these diagnostic plots might be quite satisfied proceeding with their study.

The pairs of panels (B,E) and (C,F) depict diagnostic plots for a pair of unfavorable scenarios. In (B,E), the overlap between treated and control individuals is much worse – some individuals in the dataset have a much higher probability of treatment (close to 1) than others, indicating that the randomization assumption of strongly ignorable treatment assignment may be violated. A researcher might be wary in such a situation, but might consider proceeding with a study, perhaps with some amount of trimming on the propensity score spectrum to improve propensity score overlap. (C, F) depicts an even worse possibility: the problems of propensity score overlap are the same, but prognosis and treatment assignment are also highly correllated. In a clinical study, this might mean that only the sickest patients are ever given the treatment. In a nutritional study, this might mean that only the patients with the most excellent baseline health sign up for a diet of program of interest. This is an even more serious problem because the treated individuals are systematically very divergent from the control individuals in terms of baseline variation predictive of both the potential outcome and the treatment assginment. A researcher in this situation must be very wary. Importantly, just looking at the propensity score histograms (E and F) does not allow a researcher to differentiate between these two cases. Only by visualizing the joint distribution of propensity and prognosis does the issue of highly correllated treatment assignment and prognosis come to light.

4.2 Assignment-Control Plots and Matching

Assignment-Control Plots can also be a useful diagnostic tool for matching studies. The four panels in figure 2 show the assignment-control visualizations of four different 1-to-1 matching schemes on the same data set: Mahalanobis distance (A), propensity score (B), Mahalanobis distance with a propensity score caliper (C), and Mahalanobis distance with both a propensity score caliper and a pronostic score caliper (D).

In Mahalanobis distance matching, all covariates are weighted equally in a statistical sense. When there is an abundance of uninformative covariates (i.e. those which are neither associated with the outcome nor the treatment assignment.), Mahalanobis distance matching can select matches that may actually be quite distant in the assignment-control space (Aikens, Greaves, and Baiocchi 2020). On the other hand, propensity score matching optimizes directly for matches which are nearby in terms of the variation associated with the treatment (the "assignment" axis), but it is entirely agnostic to variation associated with the outcome. This can result in high variance in estimated treatment effect (King and Nielsen 2016). Finally, the two caliper methods impose contraints on the matching option to ensure that matches are close in terms of propensity score (C) or both propensity and prognostic score (D). This visualization illustrates the potential of these methods for minimizing bias (stemming from poor propensity score balance) and variace (stemming from poor prognostic score balance) in the treatment effect.

4.3 Randomization-Assignment-Control Plots

Bhattacharya and Vogt (2007) suggest that measured covariates that are valid instruments should *not* be included in propensity score models, since this may actually *increase* the bias and variance of the causal effect estimates in the absence of strong ignorability. Intuitively, if some covariate captures some aspect of randomization or outcome-independent encouragement, we would actually like our treatment and control observations to be *far apart* in terms of the encouragement they recieve (Baiocchi et al. 2012). This supports the idea that subjects in an observational study might be best described by two separate quantities summarizing the baseline variation associated with the assignment mechanism: one "randomization" axis summarizing variation associated with the treatment assignment but unrelated to the outcome (i.e. a continuous instrumental variable or a combination of instruments), and one "assignment" axis summarizing

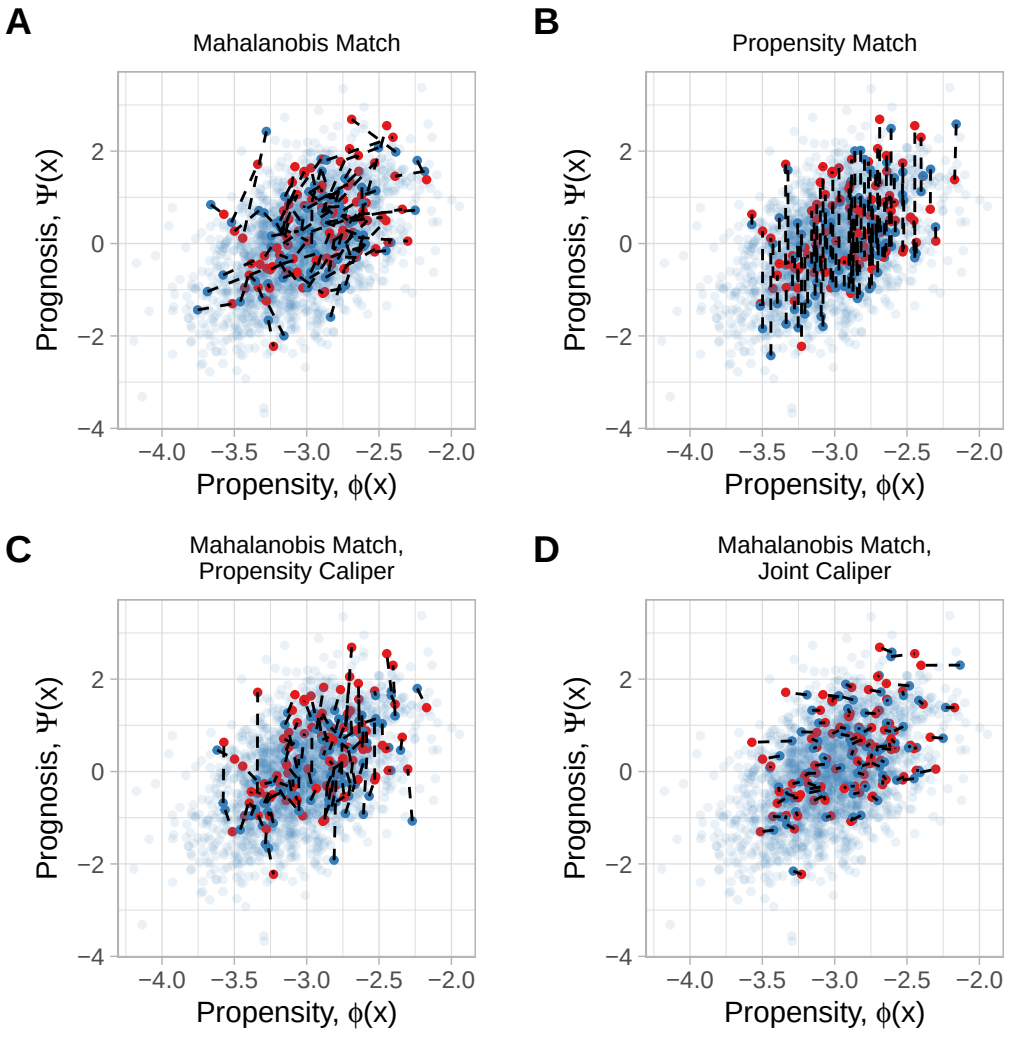

Figure 2: Assignment control plots depicting four different 1-to-1 matching schemes. Red points represent treated observations, blue points represent control. Dotted lines connect matched individuals. (A) Mahalanobis distance matching, (B), propensity score matching, (C) Mahalanobis distance matching with a propensity score caliper, (D) Mahalanobis distance matching with propensity and prognostic score calipers.

variation associated with both the treatment assignment and the outcome (i.e. the propensity score). This separation is important because observational study designs might do well to treat these two aspects of baseline variation differently. In particular, while matched subjects should ideally be very close in their propensity scores, they should ideally be very *distant* in terms of instrumental variation (Baiocchi et al. 2012).

As an illustrative example, we consider the simulation set-up as before, except that now a new measured covariate X_3, is present as an instrumental variable (IV):

$$\phi(X_i) = c_1 X_{i1} + c_2 X_{i3} - c_0,$$
$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2}.$$

Now, the ideal propensity score, $\tilde{\phi}$ would summarize just the variation in X_1 and omit any variation in X_3. Instead, X_3 should be summarized in it's own "randomization" axis. Figure 3 depicts the same matched dataset, projected down onto each pair of axes ($\tilde{\phi} \times \Psi$, $IV \times \Psi$, and $IV \times \tilde{\phi}$). The matchings shown were produced from Mahalanobis distance matching on all covariates except X_3 with propensity and prognostic score calipers. In figure 3A we see that the selected pairs are very close in terms of propensity and prognosis, but they tend to be more distant in terms of X_3 (Figure 3B-C). This visualization may be an especially useful companion to study designs using a nearfar matching approach (Baiocchi et al. 2012). The randomization-assignment-control plot may also have analouges in the randomized experiment setting, wherein the propensity score axis is omitted, and the "instrument" axis is replaced by an axis summarizing compliance behavior (e.g. "dose").

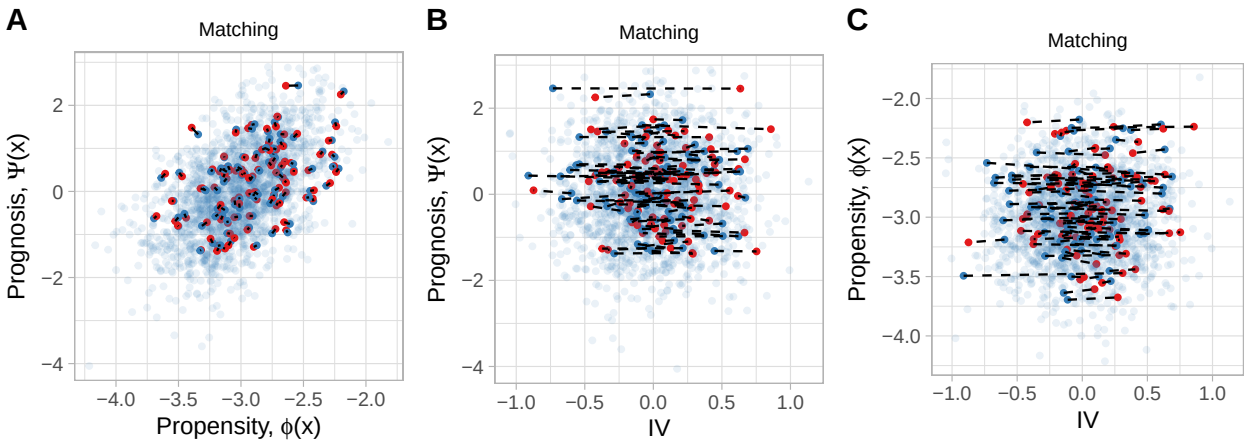

Figure 3: Randomization-Assignment-Control plots. Each panel depicts a different 2-D projection of the same dataset within the randomization-assignment-control space. Red points represent treated observations, blue points represent control. Dotted lines connect matched individuals. The matches depicted are produced by Mahalanobis distance matching on all variables except the IV with propensity score and prognostic score calipers.

A final observation to note is that the propensity and prognosis scores are correllated, but the IV is uncorrellated with either of the other two scores (Figure 3). This is actually not only expected but important: the propensity score *should* summarize variation correllated with the expected potential outcome, and the IV should be uncorrellated with the expected potential outcome in order to be an IV. Thus, we hope to expect to see correllation in the $\tilde{\phi} \times \Psi$ space, but not in $IV \times \Psi$, and $IV \times \tilde{\phi}$.

4.4 Control plots with a Forcing Variable for Regression Discontinuity Designs.

I'm struggling with this. What could this plot be used for? Maaaybe checking the continuity assumption? One would hope that the prognostic score varies smoothly across the threshold for the forcing variable - but even if that were true it doesn't guarantee the continuity assumption, and when would it not be true?

Mike, any ideas? Maybe this just isn't interesting. Maybe fuzzy RD is a better place to look?

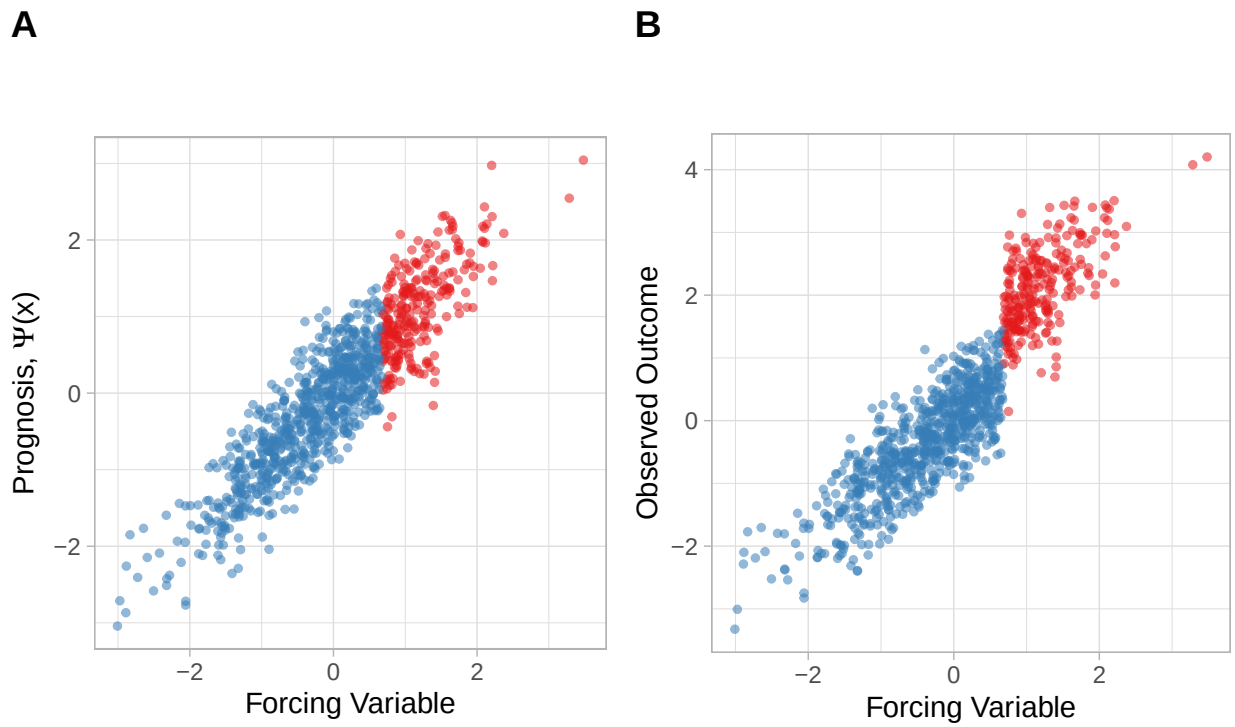

Figure 4: Diagnostic plots for a sharp regression discontinuity study. Red points represent treated subjects, blue points represent control. (A) a force-control plot, depicting each observational unit in terms of it's prognostic score and a forcing variable (B) a traditional regression discontinuity plot depicting the relationship between the outcome and the forcing variable

4.5 Assignment-Control Plots and Unmeasured Confounding

While there is a wide and increasing variety of methods which use some articulation of propensity and prognosis to estimate a treatment effect. However, the theorems underlying the use of the propensity score and the theorems underlying the use of the prognostic score both depend on the absense of unmeasured confounding. Violations of this assumption have ramifications for a wide variety of causal inference approaches, and assignment-aontrol plots may be similarly misleading when unmeasured confounding is at play.

Figure 5 illustrates the behavior of assignment control plots in a scenario with unobserved confounding. We add to our data-generating set-up an unobserved confounder, U, such that:

$$\phi(X_i) = c_1 X_{i1} - \eta U - c_0,$$
$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2} + \eta U.$$

Suppose we somehow ascertained exactly the correct relationships between the two score models and the *observed* covariates, so that our propensity and prognostic models are precisely $\hat{\phi}(X_i) = X_{i1}/c_1 - c_0$ and

$\hat{\Psi}(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2}$, respectively. That is, the score models are exactly correct, except that they do not include the unobserved confounder. Figure 5 panels A-C depict the assignment control plots we might make and matchings we might produce using these score models, $\hat{\phi}$ and $\hat{\Psi}$. Since both the assignment-control plots and the matchings use the score models $\hat{\phi}$ and $\hat{\Psi}$ that fail to capture unobserved confounding, our propensity matches appear quite close in $\hat{\phi}$ (Figure 5B) and our mahalanobis distance matchings with propensity and prognostic score calipers appear quite close in the assignment-control space defined by $\hat{\phi} \times \hat{\Psi}$ (Figure 5C).

However, panels D-F in Figure 5 show the same matches in the *true* assignment-control space, in which ϕ and Ψ are known to depend on the unobserved confounder, U. In each matching, pairs tend to differ from each other due to baseline variations in the unobserved confounder which were not accounted for in the matching process. The contrast between Figures 5C and 5F most cleanly illustrate how failing to account for U results in systematic error: in the true assignment-control space, one matched individual in each pair tends to have both higher prognostic score and higher propensity score than its partner. Since this individual is more often the treated individual than the control individual, estimates of treatment effect based on this matching will tend to be biased, since U induces systematic differences between paired individuals, even after matching.

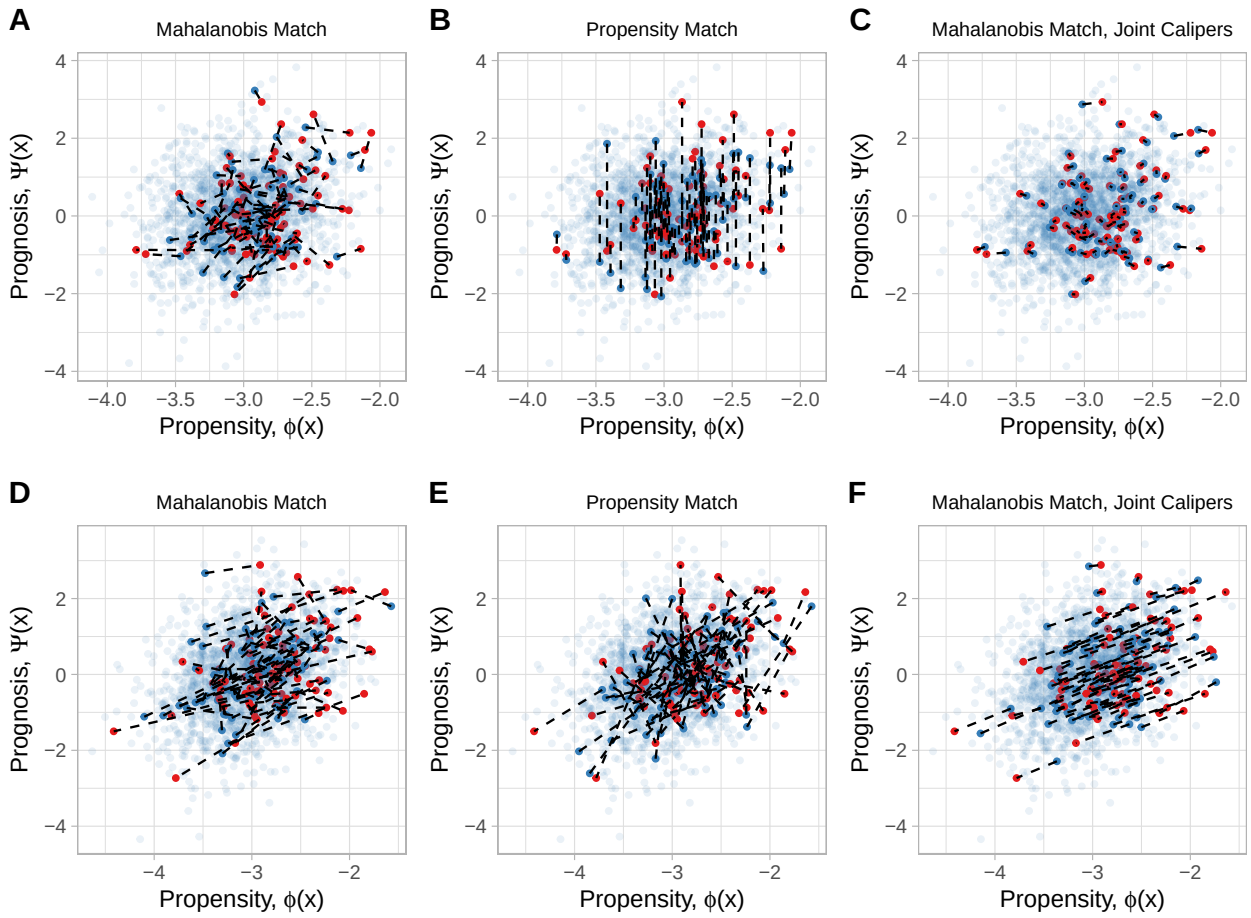

Figure 5: Assignment-control plots for three matching schemes on a dataset with unobserved confounding. A-C depict the assignment-control space as ascertained without knowledge of the unobserved confounder. D-F depict the true assignment-control space and the true match distances.

5 Conclusion

A modern shift towards an emphasis on large, passively collected datasets presents a host of challenges and opportunities for researchers interested in causality. As we collect wider datasets with more measured covariates, it will be increasingly important to prioritize the baseline variation that is most important to the causal question - correctly leveraging measured covariates which are useful, and ignoring measured covariates which are uninformative. The complementary tools of the propensity and prognostic scores are well-suited to aid in this endeavor because they summarize two important aspects of baseline variation in the measured covariates: variation associated with the assignment mechanism, and variation associated with the potential outcomes. However, they are not the only important sources of variation in a causal inference study. Other study designs may depend on an instrumental variable, a forcing variable for regression discontinuity, a summary of baseline variation associated with treatment effect heterogeneity, or compliance information in an encouragement design. Assignment-control plots and variations thereof can be thought of as dimensionality reduction tools in that they digest a possibly very large covariate space into a meaningful reduced space that is easier to use and understand. While the possible variations on this concept are numerous, they are fundamentally driven by the same concept: a principled understanding of the different types of baseline variation and their differing significances to a causal question can enable researchers to improve the design of causal inference studies and interrogate their assumptions about the data.

References

Abadie, Alberto, Matthew M Chingos, and Martin R West. 2018. "Endogenous Stratification in Randomized Experiments." *Review of Economics and Statistics* 100 (4). MIT Press: 567–80.

Aikens, Rachael C, Dylan Greaves, and Michael Baiocchi. 2020. "A Pilot Design for Observational Studies: Using Abundant Data Thoughtfully." *Statistics in Medicine*. Wiley Online Library.

Aikens, Rachael C, Joseph Rigdon, Justin Lee, Michael Baiocchi, Andrew B Goldstone, Peter Chiu, Y Joseph Woo, and Jonathan H Chen. 2020. "Stratified Pilot Matching in R: The Stratamatch Package." *Statistics arXiv*, January. https://arxiv.org/abs/2001.02775.

Antonelli, Joseph, Matthew Cefalu, Nathan Palmer, and Denis Agniel. 2018. "Doubly Robust Matching Estimators for High Dimensional Confounding Adjustment." *Biometrics* 74 (4). Wiley Online Library: 1171–9.

Baiocchi, Mike, Dylan S Small, Lin Yang, Daniel Polsky, and Peter W Groeneveld. 2012. "Near/Far Matching: A Study Design Approach to Instrumental Variables." *Health Services and Outcomes Research Methodology* 12 (4). Springer: 237–53.

Bhattacharya, Jay, and William B Vogt. 2007. "Do Instrumental Variables Belong in Propensity Scores?" 1050 Massachusetts Avenue, Cambridge, MA 02138: National Bureau of Economic Research.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95 (2). Oxford University Press: 481–88.

King, Gary, and Richard Nielsen. 2016. "Why Propensity Scores Should Not Be Used for Matching." *Copy at Http://J. Mp/1sexgVw Download Citation BibTex Tagged XML Download Paper* 378.

Leacy, Finbarr P, and Elizabeth A Stuart. 2014. "On the Joint Use of Propensity and Prognostic Scores in Estimation of the Average Treatment Effect on the Treated: A Simulation Study." *Statistics in Medicine* 33 (20). Wiley Online Library: 3488–3508.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1). Oxford University Press: 41–55.