PA

Why Propensity Scores Should Not Be Used for Matching

Gary King[1] **and Richard Nielsen**[2]

[1] *Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: king@harvard.edu, URL: http://GaryKing.org*
[2] *Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: rnielsen@mit.edu, URL: http://www.mit.edu/~rnielsen*

Abstract

We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias. The weakness of PSM comes from its attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment. PSM is thus uniquely blind to the often large portion of imbalance that can be eliminated by approximating full blocking with other matching methods. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which, we show, increases imbalance even relative to the original data. Although these results suggest researchers replace PSM with one of the other available matching methods, propensity scores have other productive uses.

Keywords: matching, propensity score matching, coarsened exact matching, Mahalanobis distance matching, model dependence

1 Introduction

Matching is an increasingly popular method for preprocessing data to improve causal inferences in observational data (Ho *et al.* 2007; Morgan and Winship 2014). The goal of matching is to reduce imbalance in the empirical distribution of the pretreatment confounders between the treated and control groups (Stuart 2010, p. 13). Lowering imbalance reduces, or reduces the bound on, the degree of model dependence in the statistical estimation of causal effects (Ho *et al.* 2007; Imai, King, and Stuart 2008; Iacus, King, and Porro 2011), and, as a result, reduces inefficiency, and bias. The resulting process amounts to a search for a data set that might have resulted from a randomized experiment but is hidden in an observational data set. When matching can reveal this "hidden experiment", many of the problems of observational data analysis vanish.

Propensity score matching (PSM) (Rosenbaum and Rubin 1983) is the most commonly used matching method, possibly even "the most developed and popular strategy for causal analysis in observational studies" (Pearl 2009). It is used or referenced in over 141,000 scholarly articles.[1]

We show here that PSM, as it is most commonly used in practice (or with many of the refinements that have been proposed by statisticians and methodologists), increases imbalance, inefficiency, model dependence, research discretion, and statistical bias at some point in both real data and in data generated to meet the requirements of PSM theory. In fact, the more balanced the

Political Analysis (2019)

DOI: 10.1017/pan.2019.11

Corresponding author
Gary King

Edited by
Jeff Gill

Authors' note: The current version of this paper, along with a Supplementary Appendix, can be found at j.mp/PScore. We thank Alberto Abadie, Alan Dafoe, Justin Grimmer, Jens Hainmueller, Chad Hazlett, Seth Hill, Stefano Iacus, Kosuke Imai, Simon Jackman, John Londregan, Adam Meirowitz, Giuseppe Porro, Molly Roberts, Jamie Robins, Bradley Spahn, Brandon Stewart, Liz Stuart, Chris Winship, and Yiqing Xu for helpful suggestions, and Connor Jerzak, Chris Lucas, Jason Sclar for superb research assistance. We also appreciate the insights from our collaborators on a previous related project, Carter Coberley, James E. Pope, and Aaron Wells. All data necessary to replicate the results in this article are available at Nielsen and King (2019).

1 Count according to Google Scholar, accessed 3/5/2019, searching for: "propensity score" AND (matching OR matched OR match).

data, or the more balanced it becomes by pruning some observations through matching, the more likely PSM will degrade inferences—a problem we refer to as the *PSM paradox*. If one's data are so imbalanced that making valid causal inferences from it without heavy modeling assumptions is impossible, then the paradox we identify is avoidable and PSM will reduce imbalance but then the data are not very useful for causal inference by any method.

We trace the PSM paradox to the particular way propensity scores interact with matching. Thus, our results do not necessarily implicate the many other productive uses of propensity scores, such as regression adjustment (Vansteelandt and Daniel 2014), inverse weighting (Robins, Hernan, and Brumback 2000), stratification (Rosenbaum and Rubin 1984), and some uses of the propensity score within other methods (e.g. Diamond and Sekhon 2012; Imai and Ratkovic 2014). Moreover, the mathematical theorems in the literature used to justify propensity scores in general, such as in Rosenbaum and Rubin (1983), are of course correct and useful elsewhere, but we show they are not relevant to the practice of matching.

We define the neglected but essential problem of model dependence in causal inference in Section 2. Suboptimal matching leads to unnecessary imbalance, which generates model dependence, researcher discretion, and statistical bias. Section 3 then proves how successfully applied matching methods can reduce model dependence. In Section 4, we show that PSM is blind to an important source of information in observational studies because it approximates a completely randomized, rather than a more informative and powerful, fully blocked experiment. It also explains the inadequacies of the statistical theory used to justify PSM. We then show, in Section 5, that PSM's weaknesses are not merely a matter of some avoidable inefficiency. Instead, when data are well balanced either to begin with or after pruning some observations by matching, the fact that PSM is approximating the coin flips of a completely randomized experiment means that it will prune observations approximately randomly, which we show increases imbalance, model dependence, and bias. As a result, other matching methods will usually achieve lower levels of imbalance than PSM, even given the same number of observations pruned, and do not generate a similar paradox until much later in the pruning process, when a fully blocked experiment is approximated and pruning is more obviously not needed.

Fortunately, since other commonly used matching methods reduce imbalance, model dependence, and bias more effectively than PSM, and do not typically suffer from the same paradox, matching in general should remain a highly recommended method of causal inference. Section 6 offers advice to those who wish to use PSM despite the problems and to those using other methods. Our Supplementary Appendix reports extensive supporting information and analyses.

2 The Problem of Model Dependence in Causal Inference

Our results apply more generally, but for expository reasons we focus on the simplest probative case. Methodologists often recommend more sophisticated approaches that encompass this simple case but, as our Supplementary Appendix demonstrates, the core intuition from the setup we give here affects these approaches in the same way, and has the advantage of being easier to understand. Thus, for unit i ($i = 1, \ldots, n$), denote the treatment variable as $T_i \in \{0, 1\}$, where 0 refers to the "control group" and 1 the "treated group". Let X_i denote a vector of k pretreatment covariates and Y_i a scalar outcome variable. In observational data, the process by which values of T are assigned is not necessarily random, controlled by the researcher, or known.

2.1 Causal Quantities of Interest

Denote $Y_i(1)$ and $Y_i(0)$ as the "potential outcomes", the values Y_i would take on if treatment or control were applied, respectively. Only one of the potential outcomes is observed for each unit i, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (Rubin 1974; Holland 1986). The treatment effect for unit i is then the difference $\text{TE}_i = Y_i(1) - Y_i(0)$.

To clarify this notation, we require two assumptions (Imbens 2004). For expository simplicity, but without loss of generality, we focus on treated units with, by definition, unobserved values of $Y(0)$. First, in order for $Y_i(0)$ and $\text{TE}_i \equiv Y_i - Y_i(0)$ to logically exist, we make the *overlap assumption*: $0 < \Pr(T_i = 0|X) < 1$ for all i (see also Heckman, Ichimura, and Todd 1998, p. 263) or, for example, that it is conceivable that any unit actually assigned treatment could have been assigned control. Second, for TE_i to be a fixed quantity to be estimated, even assuming it exists, we also assume the *stable unit treatment value assumption (SUTVA)* (Rubin 1980; VanderWeele and Hernan 2012), which requires that the potential outcomes are fixed and so, for example, the value of $Y_i(0)$ does not change if T_i, or $T_j \ \forall j \neq i$, changes from 0 to 1.

Causal quantities of interest are then averages of TE_i over different subsets of units in the sample, or the population from which we can imagine the sample was drawn. For simplicity, we focus on the sample average treatment effect (SATE), $\tau = \text{mean}_i(\text{TE}_i)$, or the sample average treatment effect on the treated (SATT), $\tau = \text{mean}_{i\in\{i|T_i=1\}}(\text{TE}_i)$ (where for set S with cardinality $\#S$, the mean over i of function $g(i)$ is $\text{mean}_{i\in S}[g(i)] = \frac{1}{\#S}\sum_{i=1}^{\#S} g(i)$).[2]

2.2 Identification

For identification, we make the *unconfoundedness assumption* (or "selection on observables", "conditional independence", or "ignorable treatment assignment"), which is that the values of the potential outcomes are determined in a manner conditionally independent of the treatment assignment: $[Y(0), Y(1)]\perp T|X$ (Barnow, Cain, and Goldberger 1980; Rosenbaum and Rubin 1983; Lechner 2001). A reasonable way to try to satisfy this assumption is to include in X any variable known to affect either Y or T, since if any subset of these variables satisfies unconfoundedness, this set will too (VanderWeele and Shpitser 2011).

Then, along with overlap and SUTVA from Section 2.1, we can identify the quantities of interest. For example, using unconfoundedness, we can identify $E[Y(0)|X = x]$ as:

$$E[Y(0)|X = x] = E[Y(0)|T = 0, X = x] = E[Y|T = 0, X = x]. \tag{1}$$

Then, extending the logic to the average identifies τ (Imbens 2004, p. 8).

2.3 Estimation Ambiguity

When feasible, we may estimate unobserved potential outcomes via exact matching. For example, we can estimate SATT with the *exact matching estimator*, $\hat{\tau} = \text{mean}_{i\in\{i|T_i=1\}}[Y_i - \hat{Y}_i(0)]$, where $\hat{Y}_i(0) = \text{mean}_{j\in\{j|X_j=X_i,T_i=1,T_j=0\}} Y_j$. Given the identification result in Equation (1), this estimator is unbiased: $E(\hat{\tau}) = \tau$.

Although exact matching is possible in hypothetical asymptotic samples, it is rarely feasible in real data sets.[3] In the common situation where exact matches are unavailable for one or more units, researchers must span the distance for each treated unit $(T_i = 1, X_i)$ to the unobserved counterfactual point $(T_i = 0, X_i)$ from the closest control units in the data set (T, X), via a statistical model, $\hat{Y}_i(0) = m_\ell(T_i = 0, X_i)$, where ℓ is an index for a model m_ℓ (part of a larger class of models defined below).

2 If some treated units have insufficiently good matches and are thereby pruned as part of the matching procedure, then the feasible SATT (or FSATT) or SATE (FSATE) may be used instead. Using FSATT or FSATE is widely recommended by statisticians (and widely used in applied research) and is appropriate so long as one is careful to characterize the resulting quantity of interest (see Crump *et al.* 2009; Rubin 2010; Iacus, King, and Porro 2011).

3 Asymptotic results are not always applicable to common practice because researchers are in the business of pushing the edge of the envelope, trying to see as far as they can given available data; if they find extra data or more certainty than they absolutely need, they do subgroup analysis, analyze interactions, or explore geographic or time series patterns that would not otherwise be discernible. Explaining the disproportionate number of t-statistics just above 2 throughout the academic literature is thus no surprise, even without considering "file drawer" problems or researchers cherry picking results. Similarly, the process of matching involves deleting observations—giving away unneeded precision for a different purpose, to reduce imbalance and bias.

The difficulty for data analysts is that different models can generate substantively different estimates of τ, even if both models fit the data well. For example, one popular choice is a linear, or weighted linear, regression of Y on T and X. Some researchers include in the regression quadratic terms or interactions for some or all of the covariates. Other popular choices include taking nonlinear transformations for Y and/or X; eliminating outliers; running robust estimators; swapping classical for one of many types of heteroskedasticity-consistent standard errors; using one of many nonlinear maximum likelihood, nonparametric, or semiparametric models; running one of the highly flexible machine learning approaches; using variable or observation selection methods; and many others. Bayesian model averaging or mixtures of expert models may help, but strong priors are usually unavailable and empirical evidence is normally insufficient to distinguish among the models.

2.4 Definition of Model Dependence

In observational data analysis, the point of the research process is to discover the data generation process rather than to design and implement one. When our knowledge of the data generation process is limited, it makes little sense to use one model as if it were known. The result of the diversity of estimates from all plausible models is that the analyst is left with model dependence—empirically different causal estimates from two or more models that fit the data approximately equally (King and Zeng 2006; Iacus, King, and Porro 2011). Levels of model dependence in real examples are often disturbingly large. Researchers respond to this ambiguity by choosing one or, at best, 4–5 results (often in different columns of a table) to publish. Crucially, the analyst chooses among the empirical estimates while selecting one result to report, which leads Ho *et al.* (2007, p. 199) to ask "How do readers know that publications are not merely demonstrations that it is *possible* to find a specification that fits the author's favorite hypothesis?"

To formalize this definition, we make two assumptions. The *fit assumption* restricts the class of models to those that fit the data approximately as well or, equivalently, that give similar predictions for potential outcomes given input points near large amounts of observed data. Denote $\tilde{x}$ as a point in the center of the data or a large subset. Then, for two models m_j and m_k $(j \neq k)$, $|m_j(\tilde{x}) - m_k(\tilde{x})| \leq h$, given a small positive constant h. In other words, the fit assumption requires that different models give similar predictions when predicting points near the data.

Second, is the *correspondence assumption*, which restricts the class of models to those which, when predicting points similar in the space of the covariates, are similar in the space of potential outcomes. Denote a Lipschitz constant K_ℓ, and two k-dimensional points, x and x', each measured in the space of the theoretical support of X (and not necessarily near its empirical support). Also define a proper nondegenerate distance, such that $d(x, x') = 0$ for exact matching (i.e., where $x = x'$) and $d(x, x') > 0$ for deviations from exact matching (i.e., where $x \neq x'$). Then the correspondence assumption is $|m_\ell(x) - m_\ell(x')| \leq K_\ell \cdot d(x, x')$. Models satisfying this assumption (after conditioning on predictors) have at least a minimal level of continuity, such as having bounded derivatives (see King and Zeng 2006; Iacus, King, and Porro 2011; Kallus 2018).

We combine these assumptions in this class of competing models (Iacus, King, and Porro 2011):

$$\begin{aligned} \mathcal{M} = \{m_\ell : |m_j(\tilde{x}) - m_k(\tilde{x})| \leq h, \quad j \neq k, \qquad & \text{(fit)} \\ \text{and } |m_\ell(x) - m_\ell(x')| \leq K_\ell \cdot d(x, x') \quad & \text{(correspondence)} \end{aligned}$$

and define *model dependence*, for any two models $m_j, m_k \in \mathcal{M}_h$ in this class and some point x in the theoretical space of X, as $|m_j(x) - m_k(x)|$ (King and Zeng 2007).

2.5 Model Dependence Biases Even Unbiased Estimators

We show here how estimators that are unbiased but inefficient when applied to one model are biased in the presence of model dependence and common researcher behavior.

Human Choice Turns Model Dependence into Bias

At a minimum, model dependence creates additional often unaccounted for uncertainty (King and Zeng 2007; Athey and Imbens 2015; Efron 2014). However, a researcher choosing among a set of estimates, rather than a set of estimators, is effectively opting for a biased estimator. Indeed, model dependence can turn even a set of unbiased estimators into a severely biased estimator. Put differently, an ex ante unbiased but inefficient estimator, conditional even on a randomly generated treatment assignment that in sample is to some degree imbalanced, is an ex post biased estimator (Robins and Morgenstern 1987).

To see this, consider a set of models $m_1, \ldots, m_J$ that lead to estimators $\hat{\tau}_1, \ldots, \hat{\tau}_J$ of the causal effect τ. Suppose we have model dependence, so that in any one data set the estimates vary: $\frac{1}{J}\sum_{j=1}^{J}(\hat{\tau}_j - \bar{\hat{\tau}})^2 > 0$, where $\bar{\hat{\tau}} = \text{mean}_j(\hat{\tau}_j)$. Assume the (unrealistically optimistic) best case: that each estimator is unbiased conditional on its model (i.e., the average over repeated samples equals the true causal estimate): $E(\hat{\tau}_j|m_j) = \tau$ (for $j = 1, \ldots, J$).

Now consider a *human-in-the-loop estimator* $\hat{\tau}_0 = g(\hat{\tau}_1, \ldots, \hat{\tau}_J)$, in which a researcher chooses one of the existing J estimates to report, in part on the basis of the empirical estimates, $\hat{\tau}_1, \ldots, \hat{\tau}_J$, not merely the models which gave rise to them, where $g(\cdot)$ is any function other than a fixed weighted average. One simple, but realistic, example is when the researcher chooses the maximum among the estimates, $\hat{\tau}_0 = \max(\hat{\tau}_1, \ldots, \hat{\tau}_J)$.

Since the researcher would likely choose a different model's estimate for each randomly drawn data set, we can no longer condition on a single model in the bias calculation and must instead condition on information from the empirical estimates. As a result, the human-in-the-loop estimator is biased: $E(\hat{\tau}_0) \neq \tau$. In other words, *a human making an unconstrained qualitative choice from among a set of different unbiased estimates is a biased estimator*. This is the reason scholars who study matching uniformly recommend that Y should not be consulted during the matching process (e.g., Rubin 2008b). The reality, of course, is usually worse than this, since some of the models in the set are likely biased.

How Biased is Human Choice?

How bad is the bias likely to be in real applications? As it happens, the social-psychological literature has shown that biases are highly likely to affect qualitative choices such as these even when researchers conscientiously try to avoid them (Banaji and Greenwald 2016) (and of course trust without verification is not an appropriate assumption about human behavior for science anyway). The tendency to imperceptibly favor one's own hypotheses, or to be swayed in unanticipated directions even without strong priors, is unavoidable. People do not have easy access to their own mental processes and they have little self-evident information to use to avoid the problem (Wilson and Brekke 1994). To make matters worse, subject matter experts overestimate their ability to control their personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005). Moreover, training researchers to make better qualitative decisions based on empirical estimates when there exists little information to choose among them scientifically is unlikely to reduce bias even if taught these social-psychological results. As Kahneman (2011, p. 170) explains, in this regard, "teaching psychology is mostly a waste of time".

Scientists are no different from other human beings in this regard. Researchers have long shown that flexibility in reporting, presentation, and analytical choices routinely leads directly to biased decisions, consistent with the researcher's hypotheses (Mahoney 1977; Ioannidis 2005; Simmons, Nelson, and Simonsohn 2011). The literature makes clear that the way to avoid these biases is to remove researcher discretion as much as possible—in the present case, by reducing model dependence—rather than instituting training sessions or encouraging everyone to try harder (Wilson and Brekke 1994, p. 118).

3 Matching to Reduce Model Dependence

For applied researchers, "the goal of matching is to create a setting within which treatment effects can be estimated without making heroic parametric assumptions", (Hill 2008). The "setting" in this quote is a subset of the data, chosen by a matching method, for which assumptions are tenable and model dependence is greatly reduced.

3.1 How Successful Matching Reduces Model Dependence

In three steps, we prove that successful matching reduces model dependence. First, we define the immediate goal of matching as finding a subset of the data closer to exact matching. Deviations from exact matching are known as *imbalance*. One way to measure imbalance is the average distance from each unit X_i to the closest unit in the opposite treatment regime, $X_{j(i)}$, i.e., where $j(i) = \arg\min_{j|T_j=1-T_i} d(X_i, X_j)$. Thus, for the original data, imbalance could be measured as $I(X) = \text{mean}_{i\in\{i\}} d(X_i, X_{j(i)})$. For a particular (matched) data subset, $\mathbb{X}$, imbalance is $I(\mathbb{X})$. Matching methods reduce imbalance when successful, so that $I(\mathbb{X}) < I(X)$.

Second, we prove that the level of imbalance in a data set bounds the degree of model dependence in estimating SATE. Denote an estimator of SATE, constructed using model $m_j \in \mathcal{M}$, as $\hat{\tau}(m_j)$, and similarly for the treatment effect $\widehat{\text{TE}}_i(m_j)$. Then:

$$
\begin{aligned}
|\hat{\tau}(m_j) - \hat{\tau}(m_k)| &= \operatorname*{mean}_{i\in\{i\}} |\widehat{\text{TE}}_i(m_j) - \widehat{\text{TE}}_i(m_k)| \\
&\leq \operatorname*{mean}_{i\in\{i\}} |m_j(X_i) - m_k(X_i)| \\
&= \operatorname*{mean}_{i\in\{i\}} |[m_j(X_i) - m_j(X_{j(i)})] + [m_k(X_i) - m_k(X_{j(i)})] + [m_j(X_{j(i)}) - m_k(X_{j(i)})]| \\
&\leq \operatorname*{mean}_{i\in\{i\}} [|m_j(X_i) - m_j(X_{j(i)})| + |m_k(X_i) - m_k(X_{j(i)})| + |m_j(X_{j(i)}) - m_k(X_{j(i)})|] \\
&\leq (K_j + K_k) \operatorname*{mean}_{i\in\{i\}} d(X_i, X_{j(i)}) + h. \qquad (2)
\end{aligned}
$$

Finally, Equation (2) implies, if matching is successful in reducing imbalance, that the bound on model dependence is lower in the matched subset than the original data:

$$
|\hat{\tau}(m_j) - \hat{\tau}(m_k)| \leq I(\mathbb{X}) < I(X) \qquad (3)
$$

which thus establishes how matching reduces the problem of model dependence.

3.2 Matching Methods

We briefly describe here PSM, and two other matching methods representative of the large variety used in the literature. We first present the simplest and most widely used version of each of the three methods and then discuss more rarely used refinements of PSM. We also report on a content analysis we conducted of the prevalence of these refinements across applied literatures.

Technical Description

Each method we define here represents one of the two existing classes of matching methods: Mahalanobis Distance Matching (MDM) is one of the longest standing matching methods that can fall within the Equal Percent Bias Reducing (EPBR) class (Rubin 1976; Rubin and Stuart 2006) and Coarsened Exact Matching (CEM) is the leading example within the Monotonic Imbalance Bounding (MIB) class (Iacus, King, and Porro 2011). PSM can also be EPBR, if used with appropriate data.

First define a function that prunes all observations from X that do not meet specified conditions: $\mathbb{X}_\ell = M(X|A_\ell, T_i = 1, T_j = 0, \delta) \equiv M(X|A_\ell) \subseteq X$, where $\mathbb{X}_\ell$ is the subset of rows of X produced by applying matching method ℓ, given condition A_ℓ. For example, under exact matching $\mathbb{X}_{\text{EM}} = M(X|X_i = X_j)$; under one-to-one exact matching with replacement $\mathbb{X}^{(1)}_{\text{EM}} = M(X|X_i = X_{j(i)})$. The nonnegative parameter δ takes on a different meaning in each matching method, where $\delta = 0$

is the best matched subset of X that can be produced according to method ℓ. Since δ is an adjustable parameter, the three methods below can be thought of as producing a sequence of matched sets, indexed by δ.

Under MDM,

$$\mathbb{X}_{\text{MDM}} = M\left(X \,\middle|\, \sqrt{(X_i - X_j)S^{-1}(X_i - X_j)} < \delta\right),$$

given a "caliper" δ (Rosenbaum and Rubin 1985a; Stuart and Rubin 2008), and sample covariance matrix S of the original data matrix X. (MDM is commonly chosen for methods articles, where the standardization makes the variables unitless; in applications, metrics such as Euclidean better enable a researcher to represent knowledge of the underlying variables and their relative importance by scaling X.)

Under CEM, $\mathbb{X}_{\text{CEM}} = M[X \mid C_\delta(X_i) = C_\delta(X_j)]$, where $C_\delta(X)$ has the same dimensions as X but with coarsened values. The parameter δ represents a chosen coarsening such that $\delta = 0$ is fine enough so that $C(X) = X$, and larger values of δ represent coarser recordings for some or all variables (larger values of δ are not necessarily ordered). Continuous covariates could be coarsened at "natural breakpoints", such as high school and college degrees in years of education, poverty level for income, etc. Discrete variables can be left as is or categories can be combined, such as when data analysts combine strong and weak Democrats into one category and strong and weak Republicans into another. (Variables can also be coarsened in groups of related variables, such as requiring the sum of three dichotomous variables to be equal.)

Finally, under PSM, $\mathbb{X}_{\text{PSM}} = M(X \mid |\hat{\pi}_i - \hat{\pi}_j| < \delta)$, where $\pi_i \equiv \Pr(T_i = 1|X_i)$ is the scalar "propensity score", in practice almost always estimated by assuming a logistic regression model $\hat{\pi}_i = (1 + e^{-X_i\hat{\beta}})^{-1}$. Most important here is the reduction of the k dimensional X_i to the scalar π_i before measuring the distance.

Content Analysis of PSM Applications

Numerous refinements of these methods, and many others, have appeared (e.g., Imbens 2004; Lunceford and Davidian 2004; Ho *et al.* 2007; Rosenbaum, Ross, and Silber 2007; Stuart 2010; Zubizarreta *et al.* 2014; Pimentel *et al.* 2018), including preceeding PSM with exact matching on a few variables and several ways of iterating between PSM and balance calculations in the space of X (e.g., Rosenbaum and Rubin 1984; Ho *et al.* 2007; Rosenbaum, Ross, and Silber 2007; Austin 2008; Caliendo and Kopeinig 2008; Stuart 2010; Imbens and Rubin 2015). The definition of $M(X|A)$ allows for matching with or without replacement, one-to-many or one-to-one matching, and optimal or greedy matching. These can be important distinctions in real data, but the issues we raise with PSM apply regardless (about which more in Section 6 and the Supplementary Appendix).

We now show that only the simplest version of PSM is used in practice with any frequency (see also Austin 2009, p. 173). To do this, we downloaded from the JSTOR repository 1,000 randomly selected English language articles, 1983–2015, which reference PSM. We then downloaded all 709 that we had permission to download with access through our university, read each one, and narrowed the list to the 279 which used PSM and, of these, the 230 which applied PSM to real data (49 additional articles were primarily methodological). We find that only 6% of the applied articles use any iterative balance checking procedure. The remaining 94% use the simplest version of PSM with one-to-one greedy matching (80%) or do so after exact matching on a few important variables, such as school district in education or age group and sex in public health (14%). We therefore use this approach in the illustrations below and give reanalyses with newer methods in the Supplementary Appendix, none of which require altered conclusions.

4 Information Ignored by Propensity Scores

Matching can be thought of as a technique for finding approximately ideal experimental data hidden within an observational data set. In three separate ways, we show here how PSM approximates an experimental design with lower standards than necessary, thus failing to use all of the information available, and generating higher levels of imbalance, model dependence, and bias.

4.1 Different Experimental Ideals for Matching Methods

Consider two experimental designs. First, under a *fully blocked randomized experimental design* (FB), such as a matched pair randomized experiment, treated and control groups are blocked at the start exactly on the observed covariates. Imbalance in these experiments is thus always 0 by design, just as what exact matching tries to accomplish after the fact, but without having to prune any observations: $X_{\text{FB}} = M(X_{\text{FB}}|X_i = X_j)$, which implies $I(X_{\text{FB}}) = 0$. Second, under a *completely randomized experimental design* (CR), treatment assignment T depends only on the scalar probability of treatment π for all units, and so is random with respect to X. In any one sample, random does not mean zero imbalance (except by rare coincidence or in asymptotic samples): $I(X_{\text{CR}}) \geqslant 0$. For simplicity, we also use the term "completely randomized" to describe partially blocked designs, such as when a constant probability of treatment is assigned to units within each of several strata, so that assignment is random, and potentially imbalanced, with respect to $(X|\pi) \neq X$.

The difference between the two experimental ideals is crucial since, compared to a completely randomized experimental design, a fully blocked randomized experimental design has more power, more efficiency, lower research costs, more robustness, less imbalance, and—most importantly from the perspective here—lower model dependence and thus less bias (Box, Hunter, and Hunter 1978; Greevy *et al.* 2004; Imai, King, and Stuart 2008; Imai, King, and Nall 2009). For example, Imai, King, and Nall (2009) found that standard errors differed in their data between the two designs by as much as a factor of six. Indeed, "for gold standard answers, complete randomization may not be good enough, except for point estimation in very large experiments", (Rubin 2008a). Of course, the discrepancy between the estimate and the truth in the one data set a researcher gets to analyze is far more important to that researcher than what happens across hypothetical repeated samples from the same hypothetical population (cf. Gu and Rosenbaum 1993).

Matching methods such as MDM and CEM approximate a fully blocked experimental design (Iacus, King, and Porro 2011, p. 349) because they come with adjustable parameters that can be set to produce the same result as exact matching, and thus zero imbalance. In particular, $\mathbb{X}_{\text{EM}} = M(X|A_{\text{CEM}}, \delta = 0) = M(X|A_{\text{MDM}}, \delta = 0)$. However, this same calculation shows that PSM approximates merely a completely randomized experiment, and thus has potentially higher imbalance. That is, because $\mathbb{X}_{\text{EM}} \subseteq M(X|A_{\text{PSM}}, \delta = 0)$, it follows that $I(\mathbb{X}_{\text{EM}}) \leq I(\mathbb{X}_{\text{PSM}})$, and strictly less than except for the unusual special cases (see also Rubin and Thomas 2000). Moreover, the fact that CEM and MDM approximate a fully blocked experiment means that each has the ability to achieve lower levels of imbalance, model dependence, and bias than PSM.

4.2 The Inadequacy of PSM Theory

The original theoretical justification given for PSM was based on the proof that unconfoundedness conditional on the raw covariates, $Y(0) \perp T \mid X$, along with overlap and SUTVA, implies unconfoundedness conditional on the scalar propensity score, $Y(0) \perp T \mid \pi$ (Rosenbaum and Rubin 1983, Theorem 1). With this result, Rosenbaum and Rubin use the identification result in Equation (1) and show that a PSM matched sample can be used to produce an unbiased estimate of SATT or SATE, conditional on one model used for estimation. The motivation for this calculation

is that it is supposedly easier to match on the scalar π than the k-dimensional X. Although this is not the case for the exact matches required by the theorem if X contains at least one continuous variable, it may be easier to find closer matches in one dimension than k.

Unfortunately, this proof, although mathematically correct, is either of little use or misleading when applied to real data. First, as Section 2.5 shows, conditioning on a single model is inappropriate, because users do no such thing. The point of observational data analysis is to discover the data generation process, and so researchers reasonably try many approaches and models. Yet, the theorem encourages researchers to settle for the lower standards of approximating only complete randomization and only average levels of imbalance (across experiments that will never be run), rather than a fully blocked experiment and balance in their own samples guaranteed to reduce model dependence. Balancing on π only is unbiased but inefficient ex ante, leaving researchers with more model dependence, discretion, and bias ex post.

The original idea behind PSM (and the proof) would have been somewhat more useful if it were reversed—if unconfoundedness (and matching) on π implied unconfoundedness on X—but this cannot be proven because it is false. Although reducing model dependence requires reducing imbalance with respect to X, balancing only on π does not balance X (since it is blind to variation in $X|\pi$). More importantly, in sample, equality between any two estimated scalar propensity scores, $\hat{\pi}_i = \hat{\pi}_j$, does not imply that the two corresponding k-dimensional covariate vectors are matched exactly, $X_i = X_j$—even though exact matching on the covariates $X_i = X_j$ does imply that the propensity scores are exactly matched $\hat{\pi}_i = \hat{\pi}_j$.

4.3 Illustration

We now simulate 1,000 data sets, each of which mixes data from three separate sources: (1) a matched pair randomized experiment, (2) a completely randomized experiment, (3) observations that, when added to the first two components, make the entire collection an imbalanced observational data set. We then study whether MDM and PSM prune individual observations in the correct order—starting with those at the highest level of imbalance (data set 3) to the lowest (data set 1). For clarity, we use two covariates (using more covariates generates patterns like those we present here, only stronger).

To fix ideas, we display one of our 1,000 data sets in the left panel of Figure 1, which highlights its three parts in separate colors. In blue in the upper right is the matched pair experiment with 25 treated units drawn uniformly from the $[-2, 2] \times [-2, 2]$ square and 25 control units which are slightly jittered versions of each of these treated units. In red, at the bottom right is a completely randomized experiment, with 50 random observations drawn uniformly from the $[-2, 2] \times [-8, -4]$ rectangle and with 25 of these randomly assigned to treatment and 25 to control. Finally, we simulate part of an imbalanced observational study by adding 50 control observations in black, drawn uniformly from the $[-6, -4] \times [-8, 2]$ square, and without corresponding randomly drawn (or otherwise overlapping) treated units.

We apply PSM and MDM, as per Section 3.2, to each data set, and iteratively remove the (next) worst observation as defined by each matching method. In the two panels at the right of Figure 1, each row of pixels stands for one simulated data set, with each individual pixel in a row representing one pruned observation color-coded by data set type. The results show that both MDM and PSM do well at removing the 50 control units that lack common support with any treated units (black is separate in both). MDM is able to separate the fully randomized experiment from the matched pair experiment (red is clearly separated from blue) but PSM is unable to separate the more informative matched pair experiment from the fully randomized experiment (red and blue are mixed).

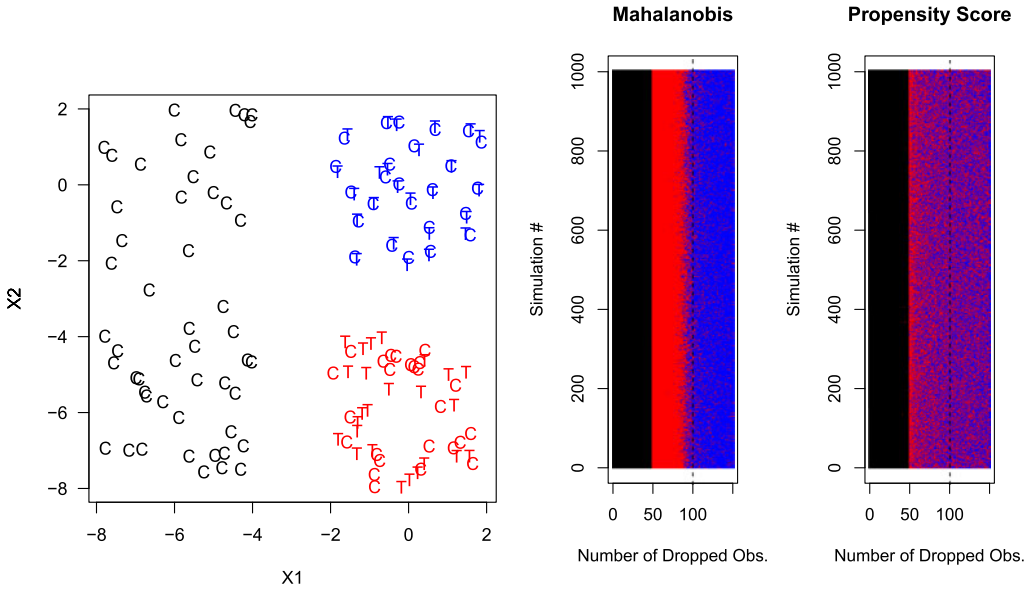

Figure 1. Finding experiments hidden in observational data, with PSM, but not MDM, blind to information from full blocking. Left panel: one (of 1,000) randomly generated data sets from a matched pair randomized experiment (in blue), a completely randomized experiment (in red), and control units from an imbalanced observational data set (in black). Right Panels: Each of the 1,000 simulations is represented by a separate row of pixels, color-coded by experiment type to indicate the order (from left to right) of which observations are pruned by MDM (left) and PSM (right).

In an application, a researcher may prefer to prune only the control units from the left part of the graph and no others. This would be best if SATT were the quantity of interest or, in some cases, to ensure that the variance is not increased too much by not pruning further. However, if the researcher chooses to prune more, and is willing to estimate FSATT, then using PSM would be a mistake. This simulation clearly shows that PSM cannot recover a matched pair experiment from these data. At best, it can recover something that looks like a fully randomized experiment, meaning that the covariates can no longer predict treatment on average. This is useful, since it makes possible estimation that is unbiased before conditioning on the treatment assignment. However, by definition some model dependence and researcher discretion remains which, when combined can lead to bias. The ideal is a fully blocked experiment, which is approximated by exact matching, not merely overlapping data clouds.

5 The Propensity Score Paradox

Given the differing goals of PSM and other methods, it is no surprise, after PSM's goal of complete randomization has been approximated, that other methods would be more effective at continuing to reduce imbalance on X than PSM. However, it also follows that pruning after this point with PSM does genuine damage—increasing imbalance, model dependence, and bias. That is, after this point, pruning the observations with the worst matched observations, according to the absolute propensity score distance in treated and control pairs, will increase imbalance, model dependence, and bias; this will also be true when pruning the pair with the next largest distance, and so on. We call this the PSM Paradox. The paradox is apparent in data designed for PSM to work well (Section 5.2) and in real applications (Section 5.3).

The reason for the PSM Paradox is because, after PSM achieves its goals of finding a subset of the data that approximates complete randomization, with approximately constant propensity scores within strata, any further pruning is at random with respect to $X|\pi$, exactly as

a completely randomized experiment. And, as we show in Section 5.1, random pruning increases imbalance.[4]

5.1 The Dangers of Random Matching

We show here that *random pruning*, a process of deleting observations in a data set independent of (T, X), not only reduces the information in the data; it also increases the level of imbalance. This may seem counterintuitive, and to our knowledge has not before been noted in the matching literature (cf. Imai, King, and Stuart 2008, p. 495). However, it is crucial, since pruning by PSM, when it succeeds in approximating complete randomization, is equivalent to random pruning. For intuition, we show this result in several ways (see also Section 1 in our Supplementary Appendix).

First, consider a completely randomized experiment with, for simplicity but no loss of generality, zero causal effects. That is, let T be generated by Bernoulli random draws and X by uniform random draws distributed over a nondegenerate space. If $k = 3$, then the expected distance of point X_i to its nearest neighbor in the opposite treatment regime $X_{j(i)}$ (among n_1 such points) is $d(X_i, X_{j(i)}) = 0.554 n_1^{-1/3}$ (Bansal and Ardell 1972). In a more general context, Abadie and Imbens (2006) show, in samples with K continuous covariates from a distribution with bounded support, that the nearest neighbor in X of a point is of order $n_1^{-1/K}$. Thus, in either framework, when a matching method prunes observations randomly, n_1 declines, the distance increases, and imbalance $I(\mathbb{X})$ grows.

Second, for intuition, consider a simple discrete data set that happened to be perfectly balanced, with a treatment group composed of one male and one female, M_1, F_1, and a control group with the same composition, M_0, F_0. Then, randomly dropping two of the four observations leaves us with one matched pair among $\{M_1, M_0\}$, $\{F_1, F_0\}$, $\{M_1, F_0\}$, or $\{F_1, M_0\}$, with equal probability. This means that with 1/2 probability the resulting data set will be balanced ($\{M_1, M_0\}$ or $\{F_1, F_0\}$) and with 1/2 probability it will be completely imbalanced ($\{M_1, F_0\}$ or $\{F_1, M_0\}$). Thus, on average in these data random matching will increase imbalance.

Finally, for a simple continuous example, consider a randomly assigned T and a fixed univariate X. Consider, as a measure of imbalance, the squared difference in means between the treated and control group of X. The expected value of this measure (which equals the squared standard error of the difference in means) is proportional to $1/n$. Thus, as we prune from this sample randomly, n declines and our measure of imbalance increases.

Of course, if all the matching discrepancies are of the same size, pruning at random or by any other means will not change the average matching discrepancy (or most other measures of imbalance). But in more realistic simulations, and real data we have examined, random pruning increases imbalance. We also introduce a higher dimensional example with real data in Section 5.3.

5.2 Simulation

We now turn to a demonstration of how PSM generates model dependence and bias. We begin by hiding a completely randomized experiment within an imbalanced data set. Unlike Figure 1, we do not include a fully blocked experiment within these data. For each of two covariates,

4 PSM is sometimes described as solving the curse of dimensionality problem (Dehejia 2004). In fact, we illustrate in Section 3 in our Supplementary Appendix that PSM's two-step procedure is an increasingly worse summary of X_i as the number of elements k in the vector increase beyond one (see Brookhart *et al.* 2006). Although the curse of dimensionality affects every matching method—and in high enough dimensions no matching method will be very effective—the problem with PSM starts with only two covariates. Other issues with PSM are well known. For one, estimating the propensity score regression with misspecification can bias estimates (Drake 1993; Smith and Todd 2005a; Kang and Schafer 2007; Zhao 2008; Diamond and Sekhon 2012). For another, if X_i contains continuous variables, π_i will be continuous and so is no easier to match exactly than X. The PSM Paradox is in addition to these important points; even when the propensity score logit is correctly specified or known, and even if matches can be found, PSM discards considerable information, and this ex ante inefficiency is equivalent to bias.

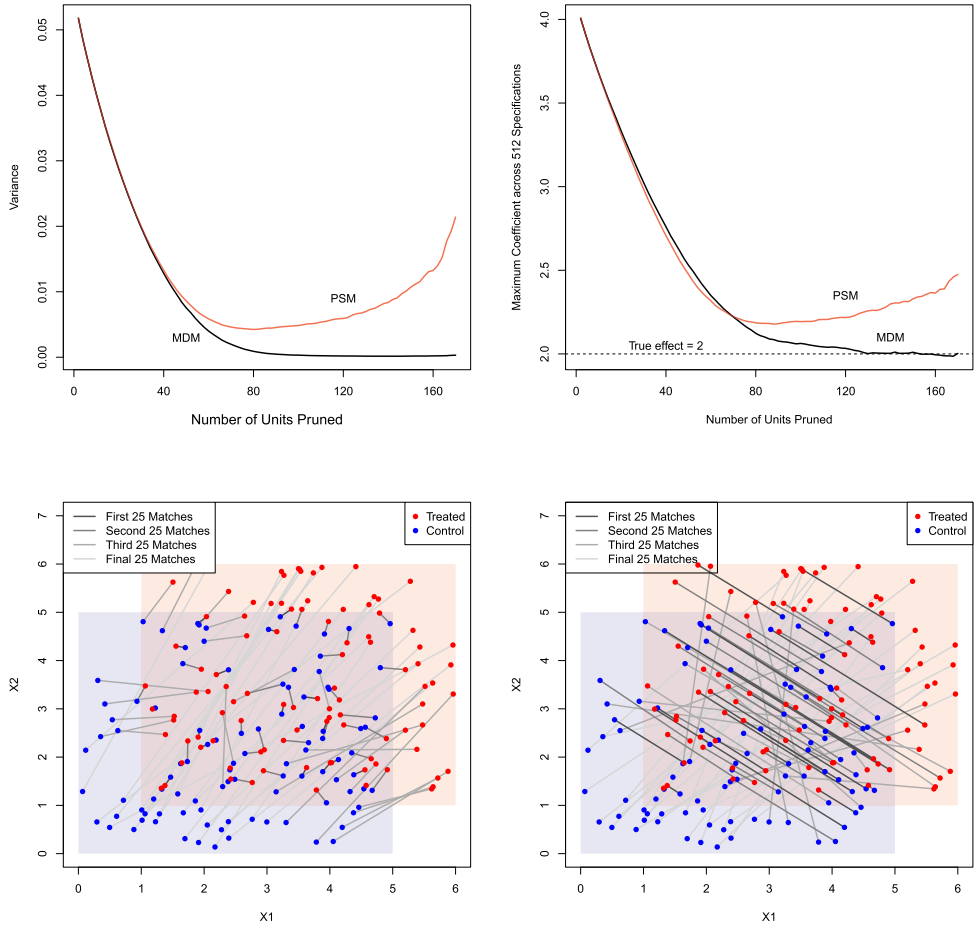

Figure 2. PSM Increases Model Dependence and Potential Bias. Top-left panel: the vertical axis measures model dependence as the average, over 100 data sets, of the variance in the causal effect estimate across 512 models applied to the same data. Top-right panel: the vertical axis shows the maximum estimated causal effect from 512 models applied to each of 100 data sets. For one simulated data set, the order of matches is indicated for MDM (bottom-left panel) and PSM (bottom-right panel).

we randomly and independently draw 100 control units from a Uniform(0, 5) and 100 treated units from Uniform(1, 6). This leaves the overlapping $[1, 5] \times [1, 5]$ square as a completely randomized experiment and observations falling outside adding imbalance. We generate the outcome as $Y_i = 2T_i + X_{i1} + X_{i2} + \epsilon_i$, where $\epsilon \sim N(0, 1)$. We repeat the entire simulation 100 times. We assume, as usual, that the analyst knows the covariates necessary to achieve unconfoundedness but does not know the functional form.

To evaluate the methods, we compute both model dependence and potential for bias, each averaged over 100 simulated data sets. We measure model dependence by the variance in the estimate of the causal effect over 512 separate models (linear regression using every possible combination of X_1 and X_2 and their 3 second order and 4 third order effects) from the same simulated data set for each given caliper level; we do this for PSM and then also MDM as a comparison. The results, which appear in the top-left panel of Figure 2, show that at first the degree of model dependence drops for both MDM and PSM, but then, after PSM has pruned enough so that the PSM paradox kicks in, model dependence dramatically increases. Instead of benefiting from units being dropped, PSM is causing damage. (This is like walking into a shoe store, giving the cashier some money and, instead of handing you a new pair of shoes, he takes the shoes you walked in with.) Model dependence for MDM, as expected, declines monotonically as stricter calipers are applied and more units are pruned.

To show how the combination of model dependence and analyst discretion can result in bias, we implemented an estimator meant to simulate the common situation where the analyst chooses a preferred estimate to publish from many possible estimates. Suppose the researcher's preferred hypothesis is that the causal effect is large, and that this preference intentionally or unintentionally affects their choice. Thus, for each caliper level of PSM and then MDM, we select the largest estimated treatment effect from among the estimates provided by the 512 possible specifications. The results, in the top-right panel of Figure 2, show that calipering initially does what we would expect by reducing the potential for bias for both MDM and PSM, with PSM even slightly outperforming MDM. However, as calipering continues, the PSM paradox kicks in, and PSM increases model dependence (as indicated in the top-left graph), the potential for bias with PSM dramatically grows even while the bias under MDM monotonically declines as we would expect and desire. (Although we do not show the graph, these patterns are unchanged for mean squared error as well.)

To provide intuition for how the paradox occurs in these data, we show which observations are matched and in which order in one of the 100 simulated data sets. Thus also in Figure 2, we plot X_1, X_2 points from one simulated data set, with matches from MDM (bottom-left panel) and PSM (bottom-right panel) denoted by lines drawn between the points, colored in by when they were matched or pruned in the calipering process. (The outcome variable is ignored during the matching process, as usual.) Darker colors were pruned later (i.e., matched earlier).

As expected, the MDM results (bottom-left panel) show that treated (circles) and control (dots) pairs that are close to each other are matched first (or pruned last). These darker blue lines mostly appear within the (completely randomized) square in the middle. In stark contrast, PSM, in the bottom-right panel, finds many matches seemingly unrelated to local closeness of treated and control units and many even outside the middle square. The diagonal pattern in PSM dark lines comes from the propensity score logit which cannot distinguish high values of X_1 and low values of X_2 from low values of X_1 and high values of X_2.

Overall, the figure shows that PSM is trying to match globally—meaning it essentially has only one chance to get it right, rather than matching locally like other methods and having some additional degree of robustness. In fact, because the propensity score is outside the space of the original data, using it for analysis violates the *congruence principle*. This principle holds that the data space and analysis space should be the same. Statistical methods which violate this principle are known to generate nonrobust and counterintuitive properties (Mielke and Berry 2007).

5.3 Damage Caused in Real Data

In this section, we reveal the PSM paradox in real applications, with data selected and analyzed by others, including two published studies and a large number of others in progress. We obtained the data from the studies in progress by advertising to assist scholars in making causal inferences, in return for access to their data and a promise not to redistribute their data or publish their substantive results (or identities). For almost all the more than 20 data sets in progress we analyzed, we found patterns similar or analogous to the two we are about to present in detail. From this, we conclude that the PSM Paradox is prevalent in many real applications.

In this first published study we reanalyze, Finkel, Horowitz, and Rojo-Mendoza (2012) showed that civic education programs in Kenya cause citizens to have more civic competence and engagement and be more supportive of the political system, with $n = 3{,}141$ survey responses, 1,347 of which received the program. They also measured a large number of socioeconomic, demographic, and leadership covariates. Second, Nielsen *et al.* (2011) show that a sudden decrease in international foreign aid to a developing country (an "aid shock") increases the probability of the onset of lethal conflict within that country. They collect data on developing countries from 1975 to 2006, in total representing $n = 2{,}627$ country-years, including 393 (treated) aid shocks.

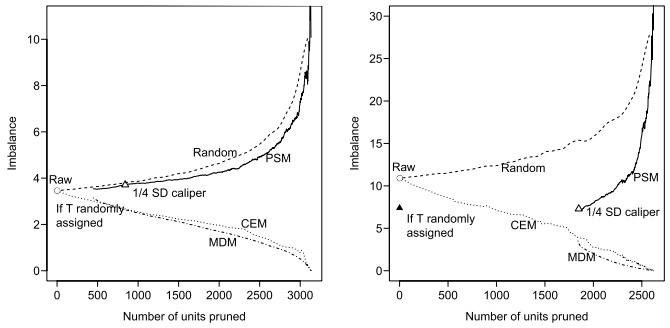

Figure 3. Imbalance-matched sample size graph, with data from Finkel, Horowitz, and Rojo-Mendoza (2012) for the left panel and Nielsen *et al.* (2011) for the right.

The authors measure 18 pretreatment covariates representing national levels of democracy, wealth, population, ethnic and religious fractionalization, and prior upheaval and violence. Finally, we analyzed a large number of data sets obtained from scholars doing work in progress, which we received by trading offers of help with their analyses and promising not to cite or scoop them. The results of all sources of data yielded very similar conclusions to that from the two data sets we now reanalyze.

For both of the published studies we replicate, Figure 3 plots imbalance (vertically) by the number of pruned observations (horizontally). We measure imbalance (i.e., the difference between the empirical distribution of X in the treated and control groups) by the "Mahalanobis Discrepancy", proposed by Abadie and Imbens (2006), which as per Section 3.1 measures imbalance in the space of X.[5] In each plot, the open circle at the left summarizes the imbalance in original data set. For reference, we also add a solid triangle that summarizes the level of imbalance that would be present if T were assigned via complete randomization.

The one-to-one PSM analysis (which in the case of Nielsen *et al.* 2011 is the published result and in both cases is estimated by all main effects in a logit model) is represented by the left end of the solid line. In the left panel, PSM's initial result is worse than the original data; in the right panel it is an improvement, as Nielsen *et al.* (2011) report. However, consider what happens in either data set if we caliper off the worst match according to the propensity score metric (i.e., the largest value of $|\hat{\pi}_i - \hat{\pi}_j|$ across all matched pairs), recalculate the level of imbalance, and repeat. These results, which are represented by the full black line in each panel, reveal the PSM paradox kicking in immediately and continuing until no data is left: That is, the more strict we are in applying PSM, the worse imbalance gets. (In a few of the unpublished data sets we analyzed that had much worse initial imbalance, PSM helped for initial pruning and then started increasing imbalance as in these graphs; simulated examples of this pattern appear in Section 3 of our Supplementary Appendix.)

If we use the venerated practice of setting the caliper to 1/4 of a standard deviation of the propensity score (Rosenbaum and Rubin 1985a), imbalance is worse than the basic PSM solution for the left panel and provides no improvement for the right panel. Following the strictures of PSM even more closely, in the hopes of finding better balance and less model dependence, accomplishes precisely the opposite.

For comparison, in each graph, we also prune via MDM (dot-dashed line) and CEM (dotted line). For MDM, we do one-to-one matching (and so the line starts at the same horizontal point as PSM) and then caliper off the observations with the largest Mahalanobis distance, recompute

5 We repeated the analysis the L_1 imbalance metric proposed by Iacus, King, and Porro 2011, and the average absolute difference in means for the columns of X, the components of which are often used in applied articles. Essentially the same conclusions result from each of these and other measures we have tried. We also repeated the analysis with various direct measures of model dependence, and found similar conclusions, although the large number of covariates in these applications mean that numerous measures could be chosen.

imbalance, and repeat. For CEM, we begin with the loosest possible coarsening, so that all data fall in a single stratum and no observations are pruned (and so the line starts from the original data). We then randomly select a variable, add one cut-point to its coarsening (always arranging the cutpoints so they divide the space between the minimum and maximum values into equal sized bins), and compute the imbalance metric. Additional cutpoints eventually lead to more observations being dropped.

As can be seen in both panels in Figure 3, the MDM and CEM lines both tend downward through their entire range with no hint of the paradox that would be represented by an upward turn like PSM: in this case, the trade-off is as it should be, in that one can reasonably choose any point along this frontier to do an analysis. (The figure also includes a dashed line marked "Random" representing the average of a sequence of data sets constructed by random pruning; as with the simpler examples in Section 5.1, the figure shows that random pruning increases imbalance.)

6 Guidance for Users

We offer guidance in this section for those accustomed to using PSM, and prefer to keep using it, and also for those willing to opt for better matching methods.

6.1 Advice for PSM Users

Our results indicate that those who wish to continue to use PSM, perhaps due to familiarity or convenience in their established work flows, would improve their work by adhering to the following points.

First, researchers using any matching method must explicitly scale variables to represent their importance (in terms of prior knowledge about effects on the outcome variable), since imbalance combines with importance to affect bias. Some claim that PSM automatically solves the problem of scaling when some variables have unknown importance, but this is untrue. Scaling is less transparent with PSM than with MDM, Euclidean distance matching, or especially CEM, but PSM users cannot avoid representing prior research in the scaling of their variables. Ignoring the issue is an arbitrary choice that may well increase bias.

Second, choosing PSM introduces avoidable risks and so researchers should report on what techniques they used to avoid the resulting problems. They should clarify how much imbalance, and therefore model dependence and bias, is left after applying PSM, especially compared to how much existed in the original data. Readers deserve to know that the researcher is not making imbalance worse due to the PSM paradox. In particular, setting a more restrictive caliper, supposedly to meet the conditions of PSM more precisely, may well increase imbalance. We recommend the routine reporting of diagnostic plots like those in Figure 3. Even carefully using PSM will likely be suboptimal compared to other matching methods, but it can sometimes be an improvement relative to the original data if used much more carefully than has been common practice (see Section 3.2).

Third, researchers should be aware that PSM can help the most in data where valid causal inferences are least likely (i.e., with high levels of imbalance) and may do the most damage in data that are well suited to making causal inferences (i.e., with low levels of imbalance). PSM is better justified (i.e., still suboptimal but not as much) when very large sample sizes are available after matching, both so as not to go past the point of complete randomization and so that the difference after matching between fully blocked and completely randomized experiments is smaller and less consequential.

Finally, researchers should understand what happens when combining PSM with other matching methods, as is sometimes recommended in the literature. One option is running PSM first, being careful to stop pruning before the PSM paradox kicks in, followed by matching directly

on X with another method. The matched sample will likely be close to only applying the second method.

A second option is to precede PSM with exact or coarsened matching on a few important covariates. The advantage of this procedure is that its first step can take one closer to full blocking than PSM alone is capable of. Its disadvantage is that it makes applying PSM more dangerous: the closer the exact matching step comes to balancing all of X, the quicker pruning with PSM will lead to the paradox and begin to increase imbalance.

A final option that has been suggested is iteratively checking balance and respecifying the propensity score regression (see Section 3.2). To the extent that these methods wind up matching more on X than the propensity score, they may sometimes perform better than PSM alone, although not as well as methods freed from the constraint of unnecessarily passing through a one dimensional propensity score. In addition to being used only rarely in the applied literature (see Section 3.2), the theoretical properties of most of these approaches have not been studied. We provide some information about this approach by replicating the PSM analyses in this paper and in our Supplementary Appendix with the automated iterative procedure proposed in Imbens and Rubin (2015, chap. 13) and several others. We find little change, and more damage the more the method relies on PSM. For one example, we replicated the top three graphs from Figure 5 in our Supplementary Appendix with the Imbens–Rubin iterative procedure and found almost imperceptible differences.

6.2 Advice for Users of Other Matching Methods

Any matching method that prunes in a manner independent of the covariates (and thus is pruning randomly) can increase imbalance. With PSM, this point, which we call the PSM paradox, kicks in after the point of complete randomization is reached, since PSM is blind to information in X not represented in the propensity score.

For other matching methods that can detect all differences in X, pruning after approximating complete randomization will continue to help reduce imbalance. Much later, after we prune enough to approximate a fully blocked experimental design, all information in X will have been exhausted. At that point, all the units are exchangeable aside from their treatment assignment and so any further pruning can only take place at random (with respect to X), which would increase imbalance. Of course, at full blocking—such as for example after exact matching—it is obvious that further pruning serves no useful purpose. We can however contrive instances where the paradox occurs with other methods, and conceivably researchers might be fooled. To illustrate, we offer two simulations in Section 4 of our Supplementary Appendix that involve pruning with MDM after using all information in X.

In the end, discarding data, which violates a basic conceptual principle of statistical inference, must be done only if you get something positive from it, such as reducing imbalance. In all applications and with all matching methods, researchers should closely study the data, the units pruned, how much imbalance and model dependence is left, and whether the process of pruning is improving or degrading inferences.

In choosing a matching method, the most important considerations are (a) ensuring one can match on all of X, so that it is at least possible to approximate a fully blocked randomized experiment, and (b) being able to encode prior knowledge about the relative importance of the variables and their combinations. For data sets with solely continuous variables, Euclidean distance matching should work well. For data with continuous, discrete, and mixed variables (such as continuous variables with natural breakpoints), CEM is the most natural; it is also considerably faster than other methods and so also works well for much larger data sets. Many other approaches aside from PSM also fit these criteria.

7 Concluding Remarks

The important insight behind PSM is to analyze an observational data set by approximating as closely as possible a completely randomized experiment. However, when feasible, approximating a fully blocked randomized experiment can be substantially better. The consequence of not doing so will in some situations merely mean that important information is left on the table—just as those who actually design experiments know to block on all available pretreatment covariates whenever feasible to avoid wasting research resources, statistical power, and efficiency. However, in the case of PSM, the problem is not merely information discarded but the damage PSM causes by continuing to prune after it has nearly accomplished its goal of approximating a completely randomized experiment; in this situation, the PSM paradox will kick in and pruning observations will also increase imbalance, model dependence, researcher discretion, and bias.

Fortunately, these problems are usually easy to avoid by switching to one of the other popular methods of matching with higher standards. However, the same paradox of matching increasing imbalance can occur with other methods when enough observations have been pruned to approximate full blocking. Although few researchers would prune observations that are exactly matched, it is important to not be fooled by problems with too few observations or matching in very high dimensional space, where no matches may exist by any method.

In any matching method, the researcher should closely follow the advice in the literature about these other methods. For example, researchers should use information about the substance of the problem being analyzed, measurement characteristics of the variables included, such as encoded in coarsenings in CEM or data measurements in MDM or Euclidean distance matching. And in all cases, researchers need to provide full information and diagnostics to readers to understand what was done.

The results described here also cast doubt on other PSM-related practices and recommendations. For example, Rubin (2008a) recommends that PSM be performed on data from completely randomized experiments; although using matching in this situation may be a good idea, using PSM can be extremely problematic because pruning would then be subject to the PSM paradox from the outset. Many have recommended that all observations more than 1/4 of a standard deviation in the propensity scores be routinely calipered off (Rosenbaum and Rubin 1985a; Rosenbaum and Rubin 1985b, p. 114; D'Augustino 1998, pp. 2269, 2271; Stuart and Rubin 2007, p. 161), but this will often result in an increase in imbalance and model dependence. Many respecify the propensity score model after removing observations without overlap, but this may make the second propensity score model less powerful and thus closer to random pruning. Most suggest including all available pretreatment variables in the propensity score logit, even those that have small or possibly nonexistent effects on the outcome (Rubin 2009, and citations in Pearl 2009), but this too may cause problems since it would make the propensity score estimates from this model closer to random, and thus generate pruning that is closer to random. The consequence of the use of PSM with all of these techniques is likely to be higher imbalance, model dependence, and bias, which is precisely what the technique was designed to avoid.

An open question worth following up is whether the PSM paradox discussed here explains some of the difficulties scholars have noticed that PSM has caused, or not solved, in real data analyses. For example, Glazerman, Levy, and Myers (2003), Smith and Todd (2005b), and Peikes, Moreno, and Orzol (2008) have each pointed to PSM requiring many more observations than they expected as one source of PSM's problems, which is the difference one would experience when running a completely randomized experiment instead of a fully blocked randomized experiment.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2019.11.

References

Abadie, A., and G. W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1):235–267.

Athey, S., and G. W. Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review Papers and Proceedings* 105(5):476–480.

Austin, P. C. 2008. "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003." *Journal of the American Statistical Association* 72:2037–2049.

Austin, P. C. 2009. "Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations." *Biometrical Journal* 51(1):171–184.

Banaji, M. R., and A. G. Greenwald. 2016. *Blindspot: Hidden Biases of Good People*. New York: Bantam.

Bansal, P. P., and A. J. Ardell. 1972. "Average Nearest-Neighbor Distances Between Uniformly Distributed Finite Particles." *Metallography* 5(2):97–111.

Barnow, B. S., G. G. Cain, and A. S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies,* vol. 5, edited by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Box, G. E. P., W. G. Hunter, and J. S. Hunter. 1978. *Statistics for Experimenters*. New York: Wiley-Interscience.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Sturmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163:1149–1156.

Caliendo, M., and S. Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22(1):31–72.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96(1):187.

D'Augustino, R. B. 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* 17:2265–2281.

Dehejia, R. 2004. "Estimating Causal Effects in Nonexpermental Studies." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, edited by A. Gelman and X.-L. Meng. New York: Wiley.

Diamond, A., and J. S. Sekhon. 2012. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95(3):932–945.

Drake, C. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effects." *Biometrics* 49:1231–1236.

Efron, B. 2014. "Estimation and Accuracy After Model Selection." *Journal of the American Statistical Association* 109(507):991–1007.

Finkel, S. E., J. Horowitz, and R. T. Rojo-Mendoza. 2012. "Civic Education and Democratic Backsliding in the Wake of Kenya's Post-2007 Election Violence." *Journal of Politics* 74(01):52–65.

Glazerman, S., D. M. Levy, and D. Myers. 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589:63–93.

Greevy, R., B. Lu, J. H. Silver, and P. R. Rosenbaum. 2004. "Optimal Multivariate Matching Before Randomization." *Biostatistics* 5(2):263–275.

Gu, X. S., and P. R. Rosenbaum. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *Journal of Computational and Graphical Statistics* 2:405–420.

Heckman, J., H. Ichimura, and P. Todd. 1998. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 65:261–294.

Hill, J. 2008. "Discussion of Research Using Propensity-Score Matching: Comments on "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003" by Peter Austin, Statistics in Medicine." *Statistics in Medicine* 27(12):2055–2061.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. URL: j.mp/matchP.

Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Iacus, S. M., G. King, and G. Porro. 2011. "Multivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106:345–361. URL: j.mp/matchMIB.

Imai, K., G. King, and C. Nall. 2009. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24(1):29–53. URL: j.mp/essrole.

Imai, K., G. King, and E. A. Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171(2):481–502. URL: j.mp/misunEO.

Imai, K., and M. Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Imbens, G. W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1):4–29.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences An Introduction*. New York: Cambridge University Press.
Ioannidis, J. P. A. 2005. "Why Most Published Research Findings are False." *PLoS Medicine* 2(8):e124.
Kahneman, D. 2011. *Thinking, Fast and Slow*. London: Macmillan.
Kallus, N. 2018. "Optimal *A Priori* Balance in The Design of Controlled Experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1):85–112.
Kang, J. D. Y., and J. L. Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4):523–539.
King, G., and L. Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159. URL: j.mp/dangerEC.
King, G., and L. Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly*, 183–210. URL: j.mp/pitfallsH.
Lechner, M. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption." In *Econometric Evaluation of Labour Market Policies*, edited by M. Lechner and F. Pfeiffer, 43–58. Heidelberg: Physica.
Lunceford, J. K., and M. Davidian. 2004. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine* 23(19):2937–2960.
Mahoney, M. J. 1977. "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System." *Cognitive Therapy and Research* 1(2):161–175.
Mielke, P., and K. Berry. 2007. *Permutation Methods: A Distance Function Approach*. New York: Springer.
Morgan, S. L., and C. Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd edn. Cambridge: Cambridge University Press.
Nielsen, R., M. Findley, Z. Davis, T. Candland, and D. Nielson. 2011. "Foreign Aid Shocks as a Cause of Violent Armed Conflict." *American Journal of Political Science* 55(2):219–232.
Nielsen, R., and G. King. 2019. "Replication Data for: Why Propensity Scores Should Not Be Used for Matching." https://doi.org/10.7910/DVN/A9LZNV, Harvard Dataverse, V1.
Pearl, J. 2009. "Myth, Confusion, and Science in Causal Analysis." Unpublished paper, http://web.cs.ucla.edu/~kaoru/r348.pdf.
Pearl, J. 2009. "The Foundations of Causal Inference." *Sociological Methodology* 40(1):75–149.
Peikes, D. N., L. Moreno, and S. M. Orzol. 2008. "Propensity Score Matching." *The American Statistician* 62(3):222–231.
Pimentel, S. D., L. C. Page, M. Lenard, and L. Keele. 2018. "Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention." *The Annals of Applied Statistics* 12(3):1479–1505.
Robins, J. M., M. A. Hernan, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.
Robins, J. M., and H. Morgenstern. 1987. "The Foundations of Confounding in Epidemiology." *Computers & Mathematics with Applications* 14(9):869–916.
Rosenbaum, P. R., R. Ross, and J. Silber. 2007. "Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association* 102(477):75–83.
Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
Rosenbaum, P. R., and D. B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:515–524.
Rosenbaum, P. R., and D. B. Rubin. 1985a. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39:33–38.
Rosenbaum, P. R., and D. B. Rubin. 1985b. "The Bias Due to Incomplete Matching." *Biometrics* 41(1):103–116.
Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.
Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63:581–592.
Rubin, D. B. 1980. "Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test", by D. Basu." *Journal of the American Statistical Association* 75:591–593.
Rubin, D. B. 2008a. "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *Journal of the American Statistical Association* 103(484):1350–1353.
Rubin, D. B. 2008b. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808–840.
Rubin, D. B. 2009. "Should Observational Studies be Designed to Allow Lack of Balance in Covariate Distributions Across Treatment Groups? *Statistics in Medicine* 28:1415–1424.
Rubin, D. B. 2010. "On the Limitations of Comparative Effectiveness Research." *Statistics in Medicine* 29(19):1991–1995.
Rubin, D. B., and E. A. Stuart. 2006. "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions." *Annals of Statistics* 34(4):1814–1826.

Rubin, D. B., and N. Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573–585.
Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–1366.
Smith, J. A., and P. E. Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics* 125(1–2):305–353.
Smith, J., and P. Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125:365–375.
Stuart, E. A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1):1–21.
Stuart, E. A., and D. B. Rubin. 2007. "Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference." In *Best Practices in Quantitative Methods*, edited by J. Osborne, 155–176. New York: Sage.
Stuart, E. A., and D. B. Rubin. 2008. "Matching with Multiple Control Groups with Adjustment for Group Differences." *Journal of Educational and Behavioral Statistics* 33(3):279–306.
Tetlock, P. E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.
VanderWeele, T. J., and M. A. Hernan. 2012. "Causal Inference Under Multiple Versions of Treatment." *Journal of Causal Inference* 1:1–20.
VanderWeele, T. J., and I. Shpitser. 2011. "A New Criterion for Confounder Selection." *Biometrics* 67(4):1406–1413.
Vansteelandt, S., and R. Daniel. 2014. "On Regression Adjustment for the Propensity Score." *Statistics in Medicine* 33(23):4053–4072.
Wilson, T. D., and N. Brekke. 1994. "Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations." *Psychological Bulletin* 116(1):117.
Zhao, Z. 2008. "Sensitivity of Propensity Score Methods to the Specifications." *Economic Letters* 98(3):309–319.
Zubizarreta, J. R., R. D. Paredes, and P. R. Rosenbaum et al. 2014. "Matching for Balance, Pairing for Heterogeneity in an Observational Study of the Effectiveness of For-Profit and Not-For-Profit High Schools in Chile." *The Annals of Applied Statistics* 8(1):204–231.

Why Propensity Scores Should Not Be Used for Matching: Supplementary Appendix

Gary King* Richard Nielsen†

January 17, 2019

Abstract

This paper is the Supplementary Appendix to Gary King and Richard Nielsen, "Why Propensity Scores Should Not Be Used For Matching," copy at j.mp/psnot

*Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138; GaryKing.org, king@harvard.edu, (617) 500-7570.

†Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139; mit.edu/∼rnielsen, rnielsen@mit.edu, (857) 998-8039.

1 PSM Approximates Random Matching

In a simple simulation, we provide intuition for how relatively balanced data makes PSM, but not MDM or CEM, highly sensitive to trivial changes in the covariates, often producing nonsensical results that approximate random matching. In the left panel of Figure 1, we generate data with 12 observations and two covariates, with one covariate plotted by the other. The data are well balanced between treated (black disks) and control (open circles) units. From these initial data, we generate 10 data sets, where we add to each of the 12 observations a small amount of random error drawn from a normal distribution with mean zero and variance 0.001. This error is so small relative to the scale of the covariates that the new points are visually indistinguishable from the original points (in fact, the graph plots all 10 sets of 12 points nearly on top of one another, but it only appears that one set is there). Next, we run CEM and MDM; in both cases, as we would expect, the treated units are matched to the nearest control in every one of the 10 data sets (as portrayed by the pair of points linked by curved solid lines).

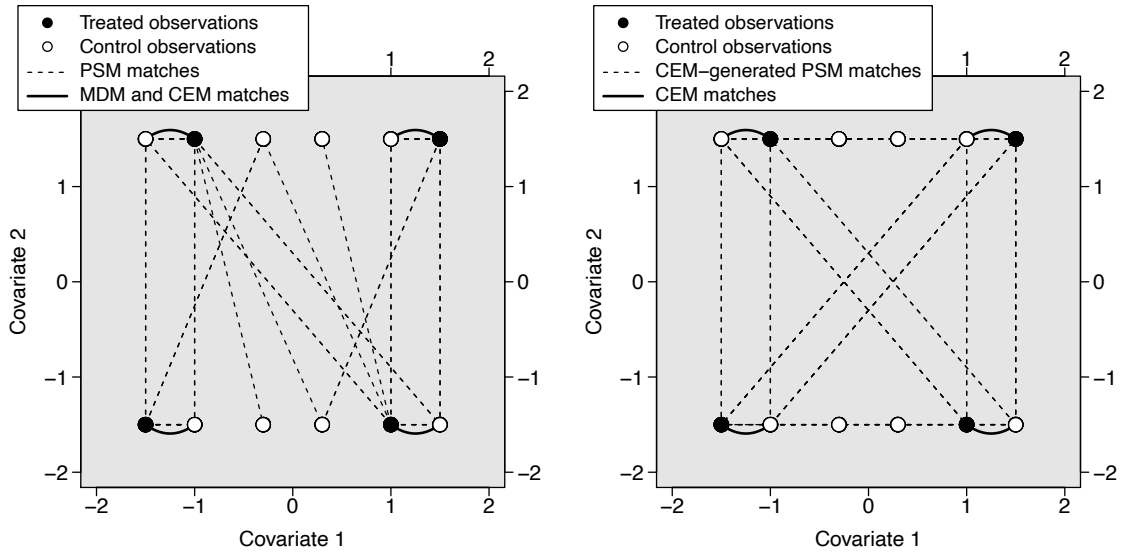

Figure 1: Ten data sets (differing from each other by imperceptibly small amounts of random error) with 4 treated units (black disks) and 8 control units (open circles). CEM and MDM match the closest control units to each treated (curved black lines). The two-step procedures match different control units for each data set, as can be seen for PSM (dashed lines, left panel) and PS-CEM (dashed lines, right panel). (The four open circles in the middle of the right panel are never matched; lines are passing through them on the way to show how other points are matched.)

However, when we run PSM on each of the 10 data sets generated for Figure 1, the four treated units are each matched to *different* control units (as portrayed by the maze of dashed lines connecting the black disks to different open circles). PSM is approximating random matching in this situation because it is unable to distinguish treated and control units; it is blind to the space of X that is not represented in $\hat{\pi}$.

Perhaps the problem is our estimation of the propensity scores? It is possible that fitting a logistic regression to twelve data points results in poorly estimated propensity scores because of finite sample bias in maximum likelihood estimators. We do not generally advocate logistic regression for twelve observations, and we only use such a small sample here for clarity in the simulation. However, the estimates of the propensity scores are not the problem. By construction, the we know the propensity scores are $0.\overline{3}$. The estimated propensity scores across all 10 simulations (120 observations) range from 0.332899 to 0.333768, so the estimation is good. Moreover, we obtain the same result if we replace the estimated propensity scores with the known propensity scores. The problem with propensity scores in this example is not about estimation.

Finally, we illustrate how the paradox results from PSM's two-step procedure. We do this by developing a (similarly unnecessary and ill-advised) two-step "propensity score CEM" (PS-CEM) algorithm: to do this, we use CEM to compute a nonparametric estimate of the propensity score (i.e., the proportion of treated units within each coarsened stratum; see Iacus, King, and Porro 2011) and, second, without running CEM as usual, we match directly on the nonparametric estimate of the propensity score. The right panel in Figure 1 is constructed the same way as the left panel except that instead of the dashed lines representing propensity score matches, they represent PS-CEM matches. The result is almost as bad as PSM. The dashed lines in the right panel show how in the different (but highly similar) data sets, the two-step PS-CEM procedure matches control units (circles) close to and also distant from treated (closed disks) units. This suggests that ignoring X and only matching based on the scalar propensity score generates the PSM paradox.

2 PSM Extensions Also Ignore Information

Most of our analysis has focused on the simplest version of PSM, which could be called *greedy nearest neighbor* matching. Our content analysis in Section 3.2 shows that the vast majority of applied papers (94%) use this simple version of PSM, but numerous extensions to PSM have been proposed in the methodological literature. We show here that these extensions to PSM do not avoid the problems we have identified. Of course, it is unsurprising that methods that seek to build on the PSM framework inherit the basic properties of and problems with PSM, even though they clearly each accomplish the more specific goals they set out to solve.

To do this, we replicate the simulation in Section 4 with six additional approaches. We first describe each briefly with citations for readers interested in further details. Two of these are adjustments to the matching procedure that introduce no new information about the covariates beyond what is contained in the propensity score:

1. **Optimal Matching** (Rosenbaum, 1989) offers an alternative to the greedy matching without replacement of the simplest version of PSM. Greedy matching without replacement matches control units to treated in order of availability, potentially resulting in poor matches for the some treated units that could have been better if considered in a different order. Optimal matching uses a network flow algorithm to construct a matched sample that minimizes the average distance between all matched pairs simultaneously.

2. **Optimal Full Matching** (Hansen, 2004; Rosenbaum, 1989) extends optimal matching from one-to-one to one-to-many groupings of treated and control units.

Three other methods incorporate additional information about the covariates along with the propensity score.

3. The **Covariate Balancing Propensity Score** (Imai and Ratkovic, 2014) approach estimates propensity scores while simultaneously optimizing covariate balance. Poorly estimated propensity scores can lead to bias. The covariate balancing propensity

score offer some robustness to poorly estimated propensity scores by using moment conditions to estimate propensity scores that are good balancing scores.

4. **Genetic Matching with the Propensity Score**. Genetic matching (Diamond and Sekhon, 2012) is a generalization of Mahalanobis distance matching that uses a genetic algorithm to estimate optimal weights for each covariate to maximize balance in the matched sample (by maximizing the p-values from paired t-tests and Kolmogorov-Smirnov tests of the covariates). When the propensity score is included as a matching covariate and receives positive weight, genetic matching is a generalization of propensity score matching as well.

5. **Mahalanobis Distance Matching with the Propensity Score as a Matching Variable**. We estimate the propensity score as normal and include it as an additional matching covariate in greedy nearest neighbor Mahalanobis distance matching. This is similar to genetic matching with the propensity score, but the contribution of the propensity score to MDM is not optimally weighted.

Finally we also include one modification of MDM with no propensity score.

6. **Genetic Matching** (ibid.) without the propensity score is a generalization of Mahalanobis distance matching that optimally weights the covariates to maximize balance (maximizing the p-values from paired t-tests and Kolmogorov-Smirnov tests of the covariates).

We also consider **PSM with a bias adjustment** proposed by Abadie and Imbens (2011), but this is an adjustment to the estimation after matching, rather than to the matching procedure itself. We suppress this from the figure because it would reveal no change at all. In any case, PSM with a bias adjustment does not do any better than simple PSM at avoiding the PSM paradox.

We use the same simulation set-up described more fully in Section 4: We simulate 1,000 data sets, each of with data from three separate sources: (1) a matched pair randomized experiment, (2) a completely randomized experiment, (3) observations that, when

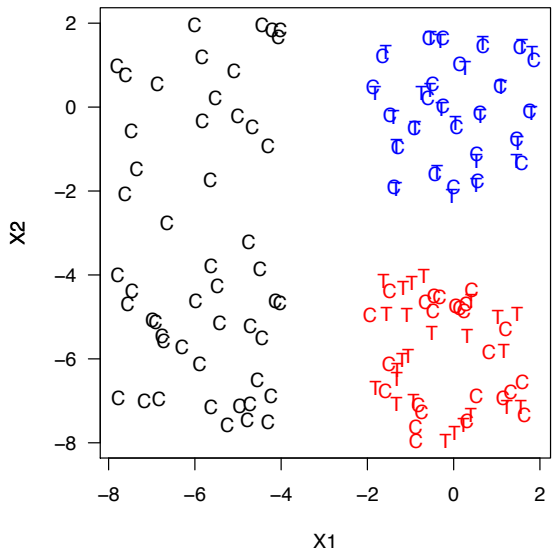

Figure 2: One (of 1,000) randomly generated data sets from a matched pair randomized experiment (in blue), a completely randomized experiment (in red), and control units from an imbalanced observational data set (in black).

added to the first two components, make the entire collection an imbalanced observational data set. Figure 2 shows one of these 1,000 data sets, with the matched pair randomized data in blue, the completely randomized data in red, and the imbalanced controls in black. Each row of pixels in the figure is a separate simulation.

We then study whether each method prunes individual observations in the correct order: starting with those at the highest level of imbalance (data set 3) to the lowest (data set 1). Ideally black is removed first, then red, then blue.

The first two panels of Figure 3 repeat the results for Mahalanobis Distance Matching and Propensity Score Matching reported in the main text. We calculate one additional statistic to aid comparison with the other variants: the proportion of observations from data set 1 that were incorrectly pruned before all observations from data sets 2 and 3 were pruned. For MDM, this proportion is 0.913 with a bootstrapped 95% confidence interval of [0.911, 0.916]. That is, approximately 8.7% of observations in data set 1 were removed too early. In comparison, the proportion of data set 1 observations correctly retained by PSM was much lower: 0.576 [0.573, 0.580]. PSM does slightly better than random guessing, but not much. Figure 4 shows these proportions graphically.

The results for the six alternative matching algorithms described above are shown in

Figures 3 and 4. Adaptations of the PSM algorithm to allow optimal matching or optimal full matching produce are no less blind to the difference between completely randomized and fully blocked data than simple PSM (see Figures 3 and 4). The covariate balancing propensity score method also offers no noticeable improvement. Matching methods based on MDM perform much better, though incorporating the propensity score hurts performance marginally, roughly in proportion to how much weight is put on the propensity score. That is, if we had 100 covariates and also the propensity score, the propensity score would not hurt as much as if there were only 2 covariates. This is as expected because the problems we identify occur with PSM and other methods that try to funnel all information about matching through the propensity score.

These results suggest that the problems we identify with PSM are not limited to the simplest cases. The blindness that PSM has to the great advantage of fully blocked data over completely randomized data is not changed even for methods that attempt to fix other, unrelated problems with PSM.

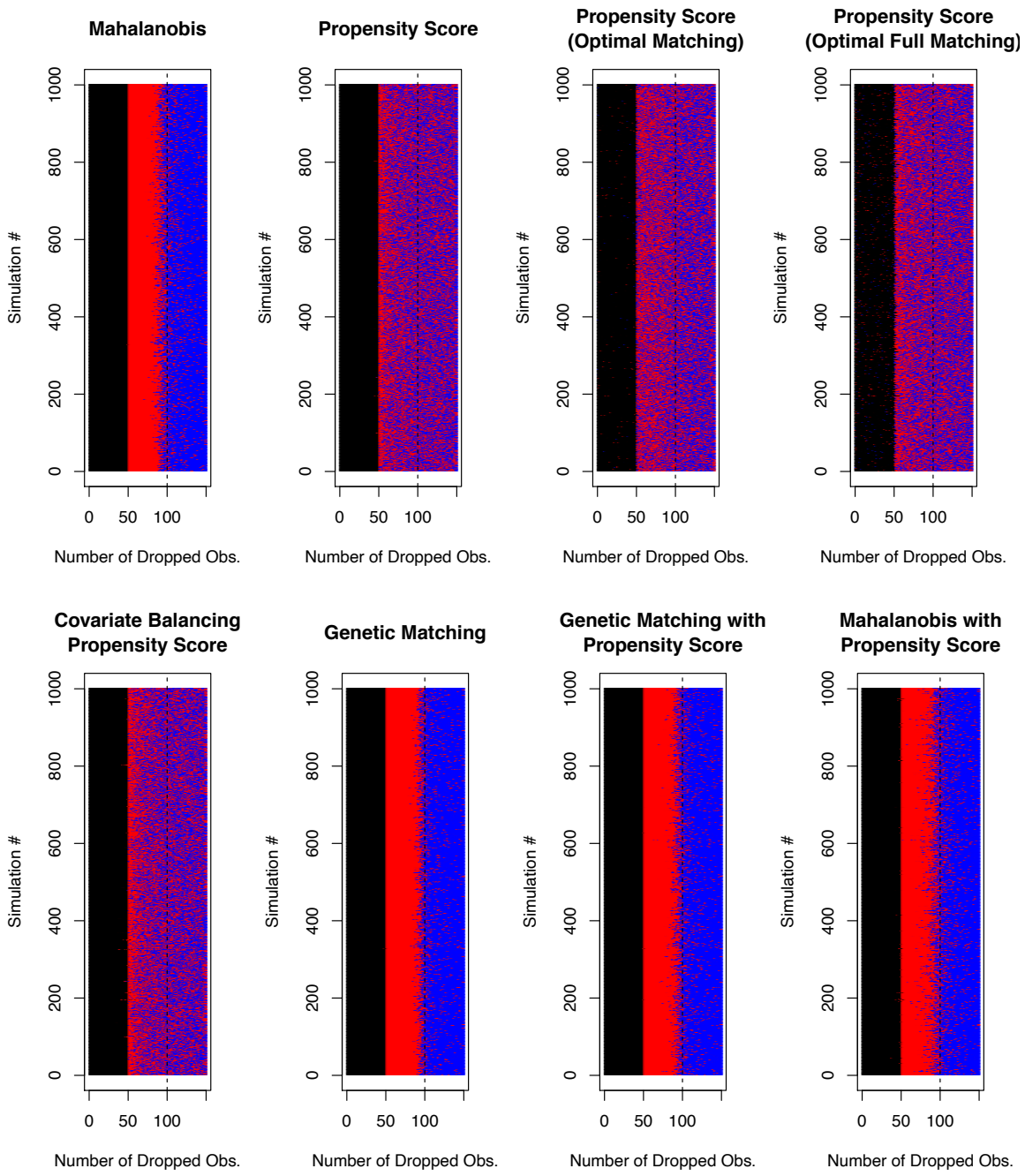

Figure 3: Variants of PSM, that build on the method, cannot find fully blocked experiments hidden in observational data. In each panel, each of 1,000 simulations is represented by a separate row of pixels, color-coded by experiment type to indicate the order (from left to right) in which observations are pruned. Ideally, black is removed first, then red, then blue. Mahalanobis, Mahalanobis with Propensity Scores, Genetic Matching, and Genetic Matching with Propensity scores generally suceed. Propensity Scores, Optimal Matching with Propensity Scores, Optimal Full matching with Propensity Scores, and the Covariate Balancing Propensity Score fail, exactly as predicted depending on how much the method depends on PSM.

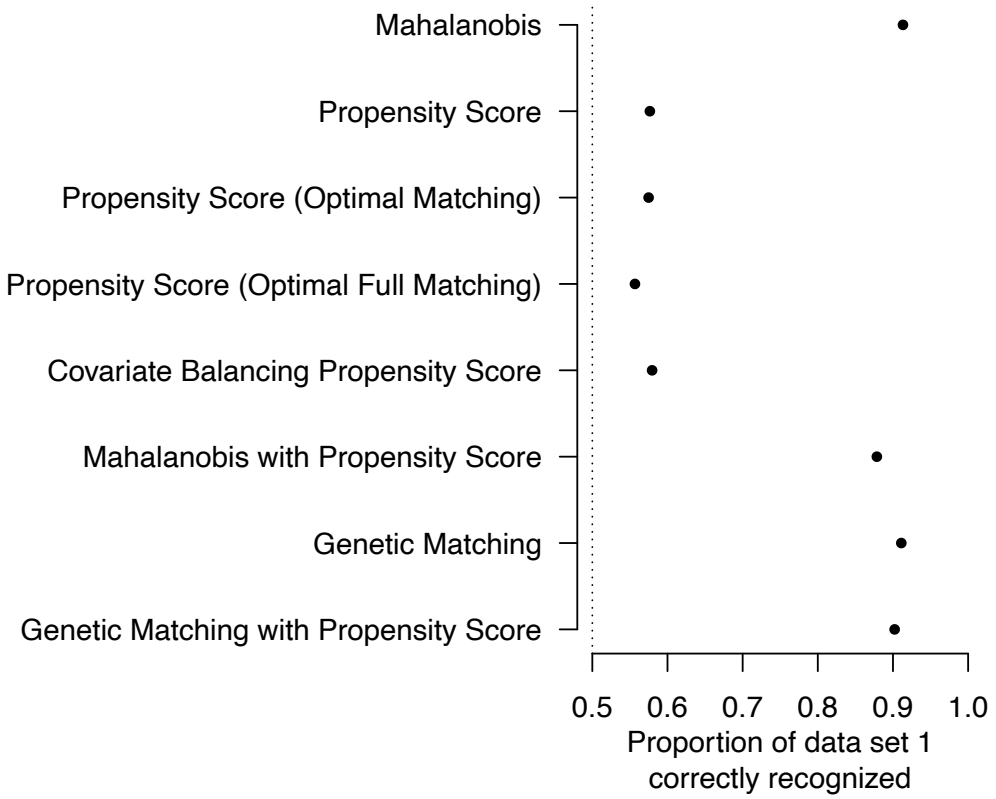

Figure 4: Comparing the performance of eight matching methods for finding fully blocked experiments hidden in observational data. The x-axis shows the average proportion of high-quality block-pair matches discarded before all non-block-pair matches are discarded by each method across 1000 simulations. Bootstrapped confidence intervals are shown, but are so small that they are difficult to see.

3 Damage Caused In Data Generated to Fit PSM Theory

We now study different types of simulations generated consistent with PSM theory, varying the number of covariates, levels of imbalance, and matching method.

3.1 Data Generation Processes

We generate data by following Gu and Rosenbaum (1993). Covariates are drawn from a multivariate normal (meeting the data requirements of EPBR) with variances of 1 and covariances of 0.2. Data sets with high, medium, and low levels of balance result from setting the control group mean vector to (0,0,0) and different treated group mean vectors to (0.1,0.1,0.1), (1,1,1), and (2,2,2), respectively. We draw 250 treated and 250 control units, which is equivalent for our purposes to generating a larger pool of controls and pruning down to 250 to achieve 1-to-1 matching with the treated units. We then prune from that point by calipering off additional units.

We draw 50 random data sets, for each of the nine combinations of 1, 2, and 3 covariates and low, medium, and high levels of imbalance. (Analyses with more covariates and higher levels of imbalance predictably produce even more dramatic patterns than presented here and so we do not present them.) For each data set generated, we analyze the same data with PSM, CEM, and MDM, following the procedures from Section 5.3. We repeat the procedure for each level of pruning and for each of the 50 data sets, average, and put a point on a graph we present below. All our results apply to estimating both SATT and FSATT; for simplicity, we present results here for the latter, which is the most commonly recommended and used approach.

For the third data generation process, we use the same simulated covariates as above, but define the treatment assignment vector using a true propensity score equation, and during estimation use the knowledge of the correct specification. We find nearly identical results for all three data generation processes, and also when rerunning them with a wide range of different parameter values; as such we only present results from the first process.

3.2 Results

Figure 5 gives the results for the methods in three rows (PSM, MDM, and CEM from top to bottom) and different numbers of covariates in separate columns (1,2,3, from left to right). Each of the nine graphs in the figure gives results from data generated with low (dotted line), medium (dashed line), and high (solid line) levels of initial imbalance. For graphical clarity, individual matching solutions do not appear and instead we average over the 50 simulations in each graph for a given matched sample size and level of imbalance. The PSM paradox is revealed whenever one of the lines increase from left to right (i.e., with imbalance increasing as the number of observations drops).

As expected, the usual curse of dimensionality reduces the performance of all three matching methods, as can be seen by the level of imbalance increasing from graphs at the left to the graphs at the right in any one row. Also as expected, the second and third rows of the Figure 5 show that CEM and MDM do not suffer from the paradox in these data: for all the lines in all the graphs in these rows, imbalance never increases as more observations are pruned, just as we would want to be the case.

However, for PSM in the first row, three important patterns emerge, all of which occur when the propensity score paradox kicks in (where a line changes direction from heading downward to where it starts heading upwards). First, no paradox emerges with PSM and one covariate (top left graph) because PSM does no dimension reduction; in this case, the propensity score is merely a rescaling of a scalar X. Second, the paradox point appears earlier, that is with fewer observations pruned, the more covariates are included in the propensity score regression (as we go from left to right in the top row of Figure 5). This problem is worse for 3 than 2 covariates and, although we do not show it, the paradox intensifies with more covariates. Third, the paradox kicks in earlier as the data become more balanced, approximating a completely randomized experiment. This can be seen by comparing the dotted (well balanced) and solid (least balanced) data in the two graphs at the top right, and noting that the point where the paradox starts moves to the left for better balanced data.

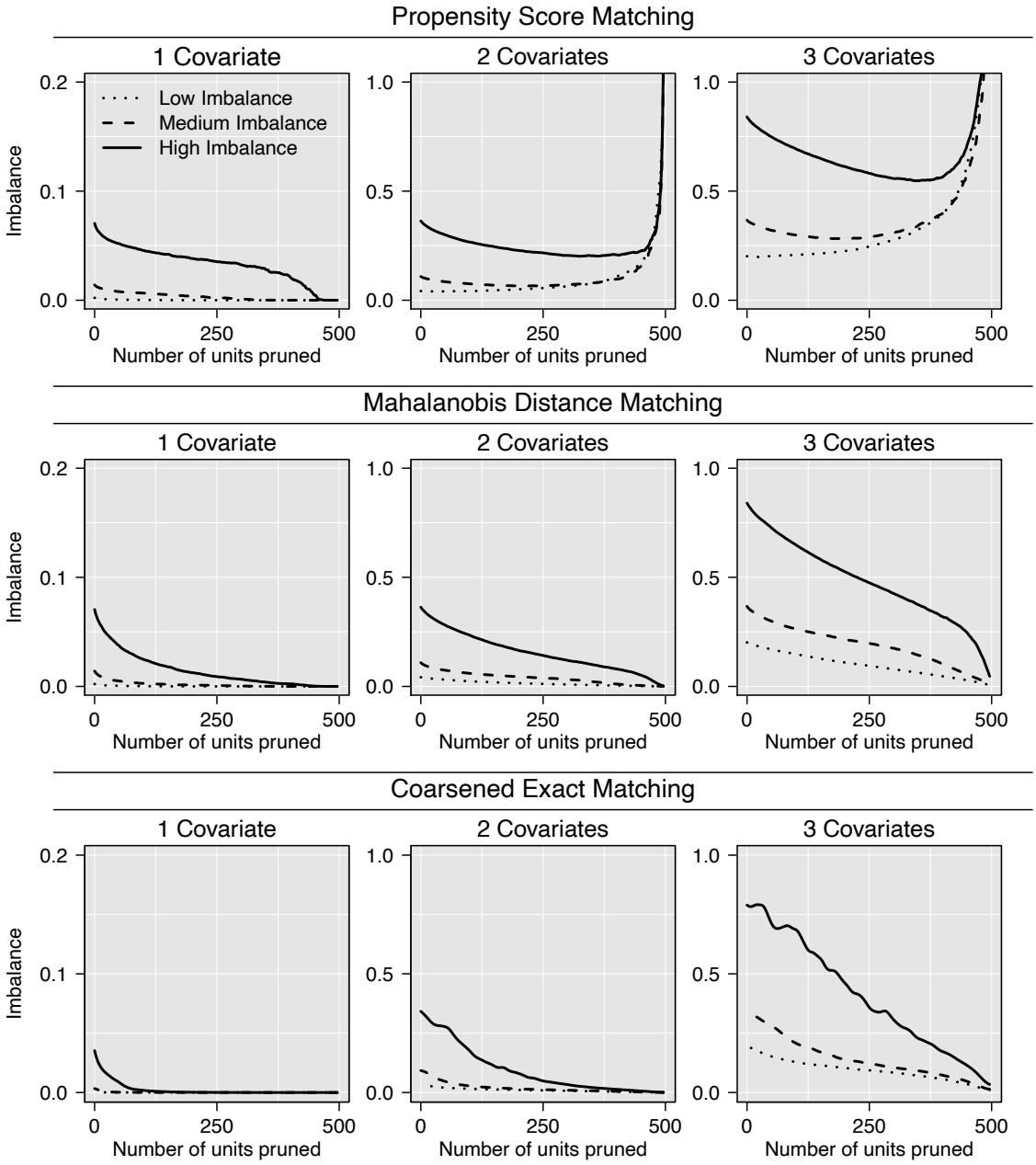

Figure 5: For PSM, MDM, and CEM in rows, and 1, 2, and 3 covariates in columns, these graphs give average values from 50 simulations with low (dotted), medium (dashed), and high (solid) levels of initial imbalance between the treated and control groups. The paradox is revealed for portions of lines that systematically increase. This can be seen for PSM with more than one covariate but not for CEM and MDM.

4 The Paradox With Other Methods

In the first simulation, we contrive a data set where nature is malicious. We begin by generating 100 values of a single covariate X deterministically, in pairs along the number line, skipping every third value, as $X = 1, 2, 4, 5, 7, 8, \ldots, 145, 146, 148, 149$. We then assign observations with even values of X to receive treatment and those with odd values to receive control. If we stopped here, each treated unit would match best to the control observation 1 unit away and T would be independent of X in sample (and where both treated and control units of X have a mean of 75). Then to each value of X, we add a tiny amount of jitter drawn from a uniform on the interval $[-0.00001, 0.00001]$. This results in some pairs being slightly better matches than others, although solely due to random jitter. We then introduce confounding (which can be productively fixed via matching) by taking the three treated units with the lowest values of X and reassigning them to control, and taking the three control units with the highest X values and assign them to treatment. For example, this creates a substantial difference between the mean value of X for the treated ($\approx$83) and control ($\approx$67). We generate the outcome variable as $Y = T + 0.01X + \epsilon$, where $\epsilon \sim N(0, 1)$.

The resulting data set has important levels of imbalance (and confounding) due to the units at the low and high values of X. The rest of the data will have matches that are effectively at random. The idea is that any method of matching will first prune the extreme (imbalanced) observations first for good reason and then start pruning at random.

We measure model dependence by first estimating the regression of Y on a constant, T, and elements of one of the subsets of $\{X, X^2, X^3, X^4, X^5\}$, and then repeat for all the other subsets. Then our measure is the range of estimates of the coefficient on T across all these regressions. Results appear in Figure 6, with model dependence plotted vertically and the number of treated units pruned by MDM horizontally.

Thus, MDM first prunes the six extreme values of X which causes model dependence to drop. After that point, when all pairs differ by pure randomness, MDM continues to prune without accomplishing anything of value. Matching in this way does not overcome the fact that pruning itself increases imbalance, and so the overall imbalance line starts

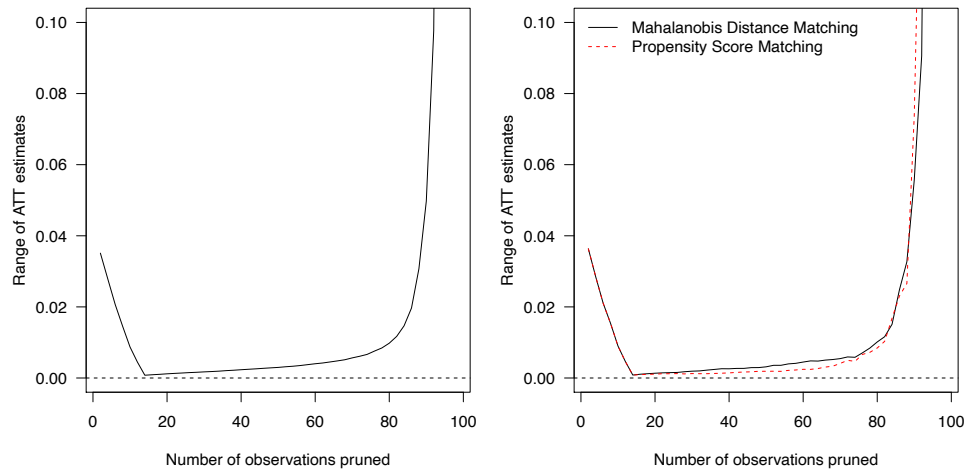

Figure 6: Left: The paradox with Mahalanobis Distance Matching. Right: Propensity Score Matching can outperform Mahalanobis Distance Matching, but only when matching is doing damage anyway.

heading upward.

We also tried to modify this simulation to create a situation where PSM outperforms MDM. However, even in this highly artificial data set, we were only able to induce better performance from PSM relative to MDM *when pruning at all was doing damage to the data set*. To do this, we increased the radius of random jitter around X from $[-0.00001, 0.00001]$ to $[-0.01, 0.01]$. As the jitter increases, PSM performs "better" — or really in this situation less worse — because it prunes at random, while MDM matches the observations with jittered X values that happen to be close first. However, MDM and PSM perform equally well at removing the six observations with extreme values of X. It is only once these six observations are removed that PSM outperforms MDM, but both methods are producing increasingly model dependent data sets at this point.

In more than two dozen real data sets and thousands of simulations we designed for this purpose, we have not seen PSM "outperform" MDM while also reducing imbalance. To be clear, we do not have a mathematical proof with such an impossibility theorem, but if it is possible it seems exceedingly unlikely in practice.

For a second illustration, we create a very small data set in high dimensional space so that points are so far spread out that few good matches are available. This is easy to see in MDM since Mahalanobis distances in this situation have the characteristic property

of differing by tiny, essentially random amounts, only after many digits to the right of the decimal point. Thus, we generate a small data set, $n = 200$, with k covariates, for $k = 2, 3, 4, 5, 10$. For each k, we generate 100 data sets with covariates drawn from independent standard normals with means drawn from a uniform on the interval $[-10, 10]$. Then, for units designated as control, we add an independent draw for each covariate from a normal with mean zero and standard deviation 5.

We then define a set $\mathcal{M}$ of linear regression models that includes all possible specifications that include subsets of covariates, squared terms, and interactions, with squared terms and interactions included only if the main effects are included. We draw one model from $\mathcal{M}$ to define the true data generating process. We use this one true model to generate Y as a linear function of the treatment times its effect of 100, the covariates with coefficients drawn from a uniform distribution on the interval of [0,500], a constant term of 500, and a normal error term with mean 0 and standard deviation 500.

For each of the 100 data sets and each sample size, we run PSM and MDM, using all main effects only. To compute model dependence for a (matched) data set, we draw 1,000 models from $\mathcal{M}$, estimate the treatment effect for each as the coefficient on the treatment variable, and then compute the variance across these estimates. In order to have a comparable measure, the subset of 1,000 models is fixed across all runs (within a fixed k). We then average the standardized estimates of model dependence within each run, over the 100 runs, and plot scaled estimates.

Figure 7 gives our results, in parallel to previous figures, so that number of units pruned is on the horizontal axis and model dependence on the vertical axis. With PSM in red and MDM in blue, one panel appears for each k. Four patterns are apparent. First, PSM has higher levels of model dependence than MDM throughout all five graphs. Second, the advantage of MDM over PSM increases in all five graphs as more observations are pruned. Third, the PSM paradox is evident in all five graphs. And finally, a paradox, with more units pruned leading to higher levels of imbalance, also affects MDM in 10 dimensional space in the last graph (and to some small extent right at the end of the some of the others).

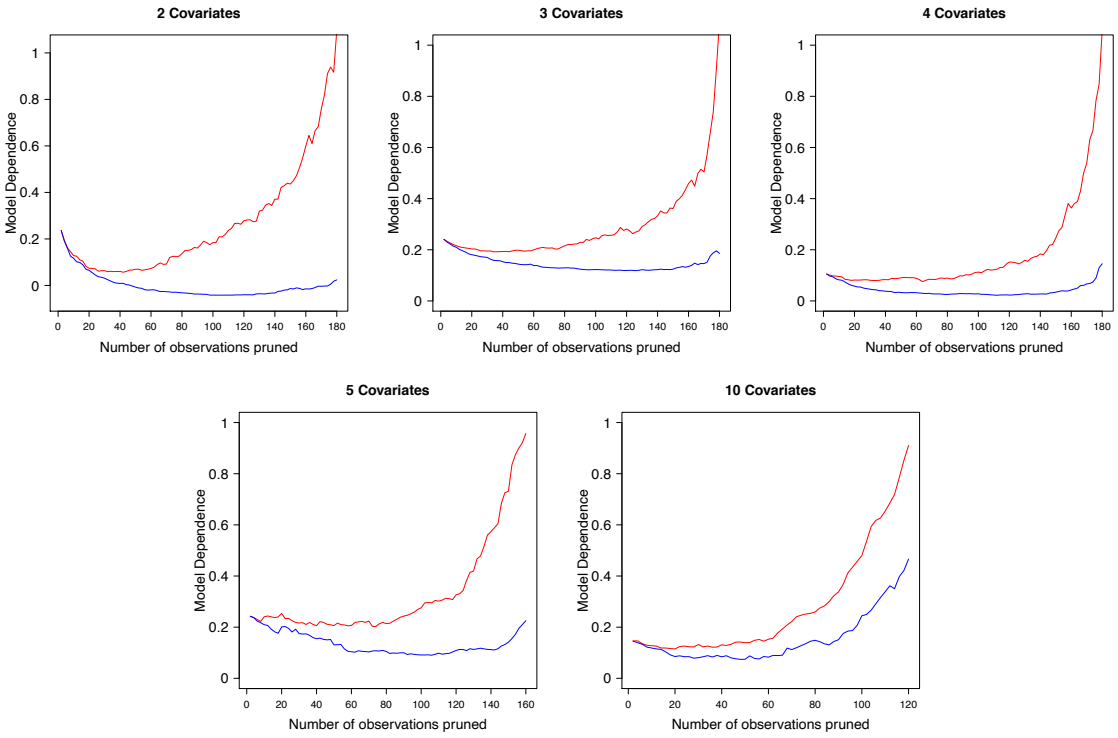

Figure 7: Model Dependence by Number of Covariates, with PSM in red and MDM in blue.

References

Abadie, Alberto and Guido W. Imbens (2011): "Bias-corrected matching estimators for average treatment effects". In: *Journal of Business & Economic Statistics*, no. 1, vol. 29.

Diamond, Alexis and Jasjeet S Sekhon (2012): "Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies". In: *Review of Economics and Statistics*, no. 3, vol. 95, pp. 932–945.

Gu, X.S. and Paul R. Rosenbaum (1993): "Comparison of multivariate matching methods: structures, distances, and algorithms". In: *Journal of Computational and Graphical Statistics*, vol. 2, pp. 405–420.

Hansen, Ben B. (2004): "Full Matching in an Observational Study of Coaching for the SAT". In: *Journal of the American Statistical Association*, no. 467, vol. 99, pp. 609–618.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2011): "Multivariate Matching Methods that are Monotonic Imbalance Bounding". In: *Journal of the American Statistical Association*, vol. 106, pp. 345–361. URL: `j.mp/matchMIB`.

Imai, Kosuke and Marc Ratkovic (2014): "Covariate balancing propensity score". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, no. 1, vol. 76, pp. 243–263.

Rosenbaum, Paul R. (1989): "Optimal matching for observational studies". In: *Journal of the American Statistical Association*, vol. 84, 1024–1032.