AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

The Balance-Sample Size Frontier in Matching Methods for Causal Inference

Gary King Harvard University
Christopher Lucas Harvard University
Richard A. Nielsen Massachusetts Institute of Technology

Abstract: *We propose a simplified approach to matching for causal inference that simultaneously optimizes balance (similarity between the treated and control groups) and matched sample size. Existing approaches either fix the matched sample size and maximize balance or fix balance and maximize sample size, leaving analysts to settle for suboptimal solutions or attempt manual optimization by iteratively tweaking their matching method and rechecking balance. To jointly maximize balance and sample size, we introduce the matching frontier, the set of matching solutions with maximum possible balance for each sample size. Rather than iterating, researchers can choose matching solutions from the frontier for analysis in one step. We derive fast algorithms that calculate the matching frontier for several commonly used balance metrics. We demonstrate this approach with analyses of the effect of sex on judging and job training programs that show how the methods we introduce can extract new knowledge from existing data sets.*

Replication Materials: The data, code, and additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: doi:10.7910/DVN/SURSEO. See also King, Lucas, and Nielsen (2016).

Matching is a statistically powerful and conceptually simple method of improving causal inferences in observational data analysis. It is especially easy to use in applications when thought of as a nonparametric preprocessing step that identifies data subsets from which causal inferences can be drawn with greatly reduced levels of model dependence (Ho et al. 2007). The popularity of this approach is increasing quickly across disciplinary areas and subfields. Indeed, the proportion of articles using matching in the *American Journal of Political Science* has increased by about 500% from 2005 to 2015, with matching used in about a quarter of all articles.

Although successful applications of matching require both reduced imbalance (increased similarity between the treated and control groups) and a sufficiently large matched sample, existing matching methods optimize with respect to only one of these two factors. Typically, the required joint optimization is performed by manually tweaking existing methods or is ignored altogether. However, jointly optimizing balance and sample size is crucial since, if the subset identified by the matching method is too small, the reduction in model dependence (and hence bias) achieved will be counterbalanced by an unacceptably high variance. Similarly, the small variance associated with a large matched data subset may be counterbalanced by unacceptably high levels of imbalance (and thus model dependence and bias). Some of these problems may also be exacerbated by current matching methods, which optimize with respect to one balance metric but encourage researchers to check the level of balance achieved with respect to a completely different metric for which the method was not designed and does not optimize.

To remedy these problems, we introduce a procedure that enables researchers to define, estimate, visualize, and then choose from what we call the *matching frontier*, a set of matched samples that fully characterize the trade-off between imbalance and the matched sample size. Unlike

Gary King is Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138 (king@harvard.edu). Christopher Lucas is Ph.D. Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138 (clucas@fas.harvard.edu). Richard A. Nielsen is Assistant Professor, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (rnielsen@mit.edu).

Our thanks to Carter Coberley, Stefano Iacus, Walter Mebane, Giuseppe Porro, Molly Roberts, Rocio Titiunik, Aaron Wells, and Yu Xie. Jeff Lewis's comments on an early version were especially helpful. Software to implement the methods proposed in this article can be found at http://projects.iq.harvard.edu/frontier.

American Journal of Political Science, Vol. 61, No. 2, April 2017, Pp. 473–489

DOI: 10.1111/ajps.12272

other approaches, we allow researchers to evaluate how much balance is achieved by pruning observations and simultaneously trade balance off against the lower variance produced by larger matched sample sizes. At each location (denoted by the matched sample size) along the matching frontier, our approach offers a matched subset of the complete data such that no other possible subset of the same size has lower imbalance. Any matching solution not on this frontier is suboptimal, meaning that a lower level of imbalance can be achieved using an alternative data subset of the same size. As a result, no matching method can outperform the matching frontier, provided that both use the same imbalance metric. Thus, constructing and analyzing a matching frontier achieves all of the benefits of any individual matching method, allows researchers to extract maximal causal information from their observational data, is considerably easier to apply appropriately than manual optimization of balance and sample size, and avoids many of the pitfalls and difficulties that lead researchers to ignore best practices in applications.

Unfortunately, finding the matching frontier for an arbitrary imbalance metric by examining each possible matched sample is an exponentially difficult computational task that is infeasible for all but the smallest data sets, even if it were possible to simultaneously use every computer ever built and run them for as long as the universe has existed. We show how to calculate the frontier for a set of commonly used imbalance metrics, for which we are able to derive a set of algorithms that run in a few minutes on one personal computer. We prove that these algorithms compute the (optimal) frontier exactly, that is, without any approximation.

We begin by introducing the trade-off between pruning observations to reduce model dependence and retaining observations to reduce variance that exists in all matching methods for causal inference. We then detail our mathematical notation, goals, and assumptions; the choices required for defining a matching frontier; and a formal definition of and algorithms for calculating the frontier. Finally, we offer several empirical examples and conclude. Software to implement all the algorithms proposed in this article can be found at http://projects.iq.harvard.edu/frontier (King, Lucas, and Nielsen 2016). Additional analyses, information, and simulations can be found in our Supplementary Appendix that accompanies this article.

The Matching Frontier Trade-Off

Matching methods selectively prune observations from a data set to reduce imbalance. A reduction in imbalance reduces, or reduces the bound on, the degree of model dependence, a result that has been shown both formally (Iacus, King, and Porro 2011a; Imai, King, and Stuart 2008; King and Zeng 2006) and in real data (Ho et al. 2007). However, matching has a potential cost in that observations pruned from the data may increase the variance of the causal effect estimate. Although researchers using matching confront the same bias-variance trade-off as in most of statistics, two issues prevent one from optimizing on this scale directly. First, since matching is commonly treated as a preprocessing step, rather than a statistical estimator, particular points on the bias-variance frontier cannot be computed without also simultaneously evaluating the estimation procedure applied to the resulting matched data set. Second, best practice in matching involves avoiding selection bias by ignoring the outcome variable while matching (Rubin 2008), which requires giving up the ability to control either bias or variance directly.

Thus, instead of bias, matching researchers focus on reducing the closely related quantity, *imbalance*. The specific mathematical relationship between the two is given by Imai, King, and Stuart (2008), but conceptually, imbalance combines with the relative importance of individual covariates to determine bias. Researchers exclude relative importance because it cannot be estimated without the outcome variable (although scaling the covariates by prior expectations of importance is a common and valuable step). Similarly, instead of variance, researchers focus on the *matched sample size*. The variance is determined by the matched sample size along with the heterogeneity (i.e., residual unit-level variance) in the data. Researchers exclude heterogeneity because it can only be estimated by using the outcome variable.

Consequently, the goal of matching involves the joint optimization of balance and matched sample size. Optimizing with respect to one, but not both, would be a mistake. Existing methods address the joint optimization by combining machine optimization of one of these factors with manual (human) optimization of the other. These attempts to optimize by hand are time-consuming and usually result in suboptimal matched samples because human data analysts are incapable of evaluating all possible matching solutions.

Many good suggestions for ad hoc approaches to manual optimization of matching methods have appeared in the methodological literature (e.g., Austin 2008; Caliendo and Kopeinig 2008; Rosenbaum, Ross, and Silber 2007; Stuart 2008). For example, Rosenbaum and Rubin (1984) detail their gradual refinement of an initial model by including and excluding covariates until they obtain a final model with 45 covariates, including seven interaction degrees of freedom and one quadratic term.

Ho et al. (2007, 216) recommend trying as many matching solutions as possible and choosing the one with the best balance. Imbens and Rubin (2009) propose running propensity score matching, checking imbalance, adjusting the specification, and iterating until convergence, as well as manual adjustments. Applying most of these methods can be inconvenient, difficult to use optimally, and hard to replicate. With nearest neighbor and propensity score matching in particular, tweaking the procedure to improve imbalance with respect to one variable will often make it worse on others, and so the iterative process can be frustrating to apply in practice.

Because manual tweaking is time-consuming and frustrating, applied researchers using matching rarely follow suggested best practices, such as the procedures just listed. In what follows, we replace the existing machine-human procedures for optimizing both balance and sample size with a machine-machine optimization procedure, thus guaranteeing optimal results in considerably less time.

Causal Inference Objectives

We define here our notation and choices for the causal quantity of interest. We separate discussion of the necessary assumptions into those that are logically part of the notation and assumptions that become necessary when trying to learn about the quantities of interest from the data.

Notation and Basic Assumptions

For unit i, let T_i denote a treatment variable coded 1 for units in the treated group and 0 in the control group. Let $Y_i(t)$ (for $t = 0, 1$) be the (potential) value the outcome variable would take if $T_i = t$. Denote the treatment effect of T on Y for unit i as $\text{TE}_i = Y_i(1) - Y_i(0)$. However, for each i, either $Y_i(1)$ or $Y_i(0)$ is observed, but never both (which is known as the fundamental problem of causal inference; Holland 1986). This means we observe $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Finally, define a vector of k pretreatment control variables X_i.

We simplify this general framework by restricting ourselves to consideration of TEs only for treated units.[1] In this situation, the only unobservables are $Y_i(0)$ for units that received treatment $T_i = 1$ (since $Y_i(1) \equiv Y_i$ is observed for these units).

[1] Since the definition of which group is labeled "treated" is arbitrary, this does not restrict us in practice.

A coherent interpretation of this notation implies two assumptions (Imbens 2004). The first is *overlap* (sometimes called "common support"): $\Pr(T_i = 1|X) < 1$ for all i (see also Heckman, Ichimura, and Todd 1998, 263). The idea here is that for treated units, where $T_i = 1$, it must be conceivable that before treatment assignment, an intervention could have taken place that would have assigned unit i instead to the control group, while holding constant the values of X. If this could not have been possible, then $Y_i(0)$, which we need to define TE_i, does not even logically exist.

A second assumption implied by the notation is *stable unit treatment value* (SUTVA), which can be thought of as logical consistency, so that each potential value is fixed even if T changes (or conceptualized differently, this assumption requires no interference between units, and no hidden versions of treatment; VanderWeele and Hernan 2012). If overlap holds but SUTVA does not, then $Y_i(0)$ and TE_i exist but are not fixed quantities to be estimated.

Quantities of Interest

From the basic definitions given above, we can compute many quantities, based on the average of TE over given subsets of units. We focus on two in this article. We first define these theoretically and then explain how they work in practice, followed by the assumptions necessary for identification.

First is the *sample average treatment effect on the treated*, $\text{SATT} = \text{mean}_{i\in\{T=1\}}(\text{TE}_i)$, which is TE averaged over the set of all treated units $\{T = 1\}$ (Imbens 2004).[2] If matching only prunes data from the control group, SATT is fixed throughout the analysis.

Second, since many real observational data sets contain some treated units without good matches, analysts often choose to compute a causal effect among only those treated observations for which good matches exist. We designate this as the *feasible sample average treatment effect on the treated* (FSATT).[3] Other possibilities include

[2] Formally, for set S with cardinality $\#S$, define the average over i of function $g(i)$ as $\text{mean}_{i\in S}[g(i)] = \frac{1}{\#S}\sum_{i=1}^{\#S} g(i)$.

[3] Using FSATT is common in the matching literature but may be seen as unusual elsewhere because the quantity of interest is defined by the statistical procedure. In fact, this approach follows the usual practice in observational data analysis of collecting data and making inferences only where it is possible to learn something. The advantage here is that the methodology makes a contribution to a step previously considered outside the statistical framework (e.g., Crump et al. 2009; Iacus, King, and Porro 2011a), just as measurement error, missing data, selection bias, and other issues once were. As Rubin (2010, 1993) puts it, "In many cases, this search for balance will reveal that there are members of each treatment arm

TE averaged over all observations, as well as population average treatment effects.

Although the distinction between SATT and FSATT is clear given a data set, the distinction can blur in practice because in observational data analysis there often exists no "correct" or canonical definition of the target population. Usually, if observational data analysts have access to more relevant data, they use it; if a portion of the data is not helpful or too difficult to use, because of measurement problems or the extreme nature of the counterfactual inferences required, they drop it. If we have a new, more powerful telescope, we observe more—for that reason. Thus, since observational data analysis is, in practice, an opportunist endeavor, we must recognize first that even the choice of SATT as a quantity of interest always involves some feasibility restriction (quite like FSATT), either explicitly where we choose to make a SATT inference in a chosen subset of our data, or implicitly due to the choice of our data set to begin with. Thus, regardless of the definition of the units over which an average will take place, researchers must always be careful to characterize the resulting new estimand, for which we offer some tools below.

Suppose we insist on choosing to estimate SATT even though some counterfactuals are so far from our data that they have no reasonable matches and require model-based extrapolation. If SATT really is the quantity of interest, this situation cannot be avoided, except when it is possible to collect more data. To understand this problem, we follow Iacus, King, and Porro (2011a) and partition the N treated units into N_f treated units that can be well matched with controls (a "feasible estimation set") and N_{nf} remaining treated units that cannot be well matched (a "nonfeasible estimation set"), such that $N = N_f + N_{nf}$. In this case, we express SATT as a weighted average of an estimator applied to each subset separately:

$$\text{SATT} = \frac{\text{FSATT} \cdot N_f + \text{NFSATT} \cdot N_{nf}}{N}. \qquad (1)$$

When estimating SATT in this unfortunate, but common, situation, it is often worthwhile to compute its two subcomponents separately since only FSATT will be estimatable without (much) model dependence. We refer to the subsets of observations corresponding to FSATT and NFSATT, respectively, as the *overlap set* and *nonoverlap set*.[4]

who are so unlike any member of the other treatment arm that they cannot serve as points of comparison for the two treatments. This is often the rule rather than the exception, and then such units must be discarded . . . . Discarding such units is the correct choice: A general answer whose estimated precision is high, but whose validity rests on unwarranted and unstated assumptions, is worse than a less precise but plausible answer to a more restricted question."

Statistical Assumptions

To establish statistical properties for estimators of these quantities, statisticians typically posit an extra layer of complication by imagining a superpopulation from which the observed data are drawn repeatedly and then attempting to infer fixed population quantities from the average over hypothetical applications of an estimator to repeated hypothetical draws from the population.[5] We first explain the assumption necessary in this hypothetical situation, simplified for SATT, but then go a step further and establish conditions under which we would get the correct answer in our one sample, without an extra layer of complication. As it turns out, these conditions are simpler and easier to understand.

First, for formal statistical identification, we make an *ignorable treatment assignment* (ITA) assumption, which for SATT requires that the mechanism that produced the treatment assignment (i.e., the values of T) be independent of the potential outcome $Y_i(0)$ given X: $T_i \perp Y_i(0) | X$ for all treated units (Barnow, Cain, and Goldberger 1980; Rosenbaum and Rubin 1983). This independence assumption can be weakened in ways that are not usually crucial distinctions in practice (Imbens 2004). (This assumption has also been referred to as "selection on observables," "unconfoundedness," and "conditional

[4] Although these definitions borrow language from the related overlap assumption introduced in the previous subsection, the two are distinct: Regardless of whether a matching counterfactual observation exists to estimate $Y_i(0)$, we need to ensure that the ex ante probability of an alternative treatment assignment would have been possible for observation i. However, if the overlap assumption is violated, it would be impossible to find a suitable counterfactual observation.

In all cases, ways of estimating the overlap set (see above) necessarily depend on substantive characteristics of the data, but methodologists usually attempt to offer some guidance on the basis of the data alone. The simplest and most stringent existing definition for the overlap region is exact matching (Manski 1995). However, in part because in most applications this definition would result in almost all observations being in the nonoverlap set and in part because reasonable smoothness assumptions make extrapolating small distances over continuous space relatively safe (Zhao 2004), most scholars choose more relaxed definitions. Some others include definitions based on nonparametric estimation of the propensity score, the quality of the worst individual matches (Imbens 2004), and the convex hull (King and Zeng 2006). In the next subsection, we offer approaches that seem naturally implied by each imbalance metric. In this way, we reduce the number of adjustable parameters to be chosen or assumed while using our methodology.

[5] Other, alternative sampling and modeling frameworks are sometimes suggested instead, but all add an extra imaginary layer of some kind.

independence"; special cases of it are referred to as "exogeneity" and "no omitted variable bias," among others.) Perhaps the simplest way to satisfy this assumption is to include in X any variable that from prior research or theory is known to cause either Y or T, since if any subset of these variables satisfies ITA, this set will too (VanderWeele and Shpitser 2011).

Second, for clarity, we pare down ITA and its superpopulation sampling framework to its essentials necessary for getting the correct answer in the sample: $\widehat{\text{SATT}} = \text{SATT}$. We retain the counterfactual framework inherent in the definition of causality so that $Y_i(0)$ is unobserved but defined for treated units. However, for point estimates, imagining infinite random draws from an invented population is unnecessary. And if we wish to consider this (e.g., for some types of uncertainty estimates), we shall follow the principle of privileging the sample in hand so that SATT is defined over the treated units in the only data set we actually observe.

The idea of matching is to replace the unobserved $Y_i(0)$ for each treated unit i with an observed $Y_j(0) \equiv Y_j$ for a control unit (i.e., $T_j = 0$) with matching covariate values ($X_i = X_j$). A sufficient (but not necessary) condition for $\widehat{\text{SATT}} = \text{SATT}$ is $(Y_i(0)|T_i = 1, X_i) = (Y_j(0)|T_j = 0, X_i)$ for all treated units i (with matching controls j). However, if any $Y_j(0)$ from individual matches does not equal $Y_i(0)$, we can still estimate SATT correctly so long as the estimates of $Y_i(0)$ are right *on average* over treated units. Note that this sounds like, but is distinct from, the concept of "unbiasedness," which refers to averages over hypothetical repeated draws from an imaginary superpopulation, rather than our need for being correct on average over the real in-sample treated units. We formalize this idea with the less restrictive *uncorrelated treatment assignment* (UTA) assumption, which is that $Y(0)$ is uncorrelated with T, within strata of X, or equivalently but more simply: $\text{mean}_i(Y_i(0)|T = 0, X) = \text{mean}_j(Y_j(0)|T = 1, X)$, which means that within strata defined by X, the average of $Y_i(0)$ for (unobserved) treated and (observed) control units is the same.

An even easier way to understand the UTA assumption is to consider the case with a data set composed of one-to-one exact matches. Exact matching is equivalent to conditioning on X, and one-to-one matching means that weighting within strata is not required (see the next section). In this simpler situation, X is irrelevant and the assumption is simply $\text{mean}(Y_i(0)|T = 1) = \text{mean}(Y_j(0)|T = 0)$, which reduces to $\text{mean}_i(Y_i(0)|T_i = 1) = \text{mean}_j(Y_j|T_j = 0)$ since the second term is fully observed.

When used in practice, applying UTA (or ITA) requires both (a) choosing and measuring the correct variables in X and (b) using an analysis method that controls sufficiently for the measured X so that T and $Y(0)$ are sufficiently unrelated that any biases that a relationship generates can be ignored (e.g., such as if they are much smaller than the size of the quantities being estimated). Virtually all observational data analysis approaches, including matching and modeling methods, assume (a) holds as a result of choices by the investigator. This includes defining which variables are included in X and ensuring that the definition of each of the variables has a meaningful scaling or metric. Then, given the choice of X, the methods distinguish themselves by how they implement approximations to (b) given the choice for the definition of X.

Matching Frontier Components

The matching frontier (defined formally in the next section) requires the choice of options for four separate components. In addition to the quantity of interest, they include and we now describe fixed- versus variable-ratio matching, a definition for the units to be dropped, and the imbalance metric.

Fixed- or Variable-Ratio Matching

Some matching methods allow the ratio of treated to control units to vary, whereas others restrict them to have a fixed ratio throughout a matched data set. Matching with replacement usually generates variable-ratio matching. Examples of fixed-ratio matching include simple 1-to-1, 1-to-p (for integer $p \geq 1$), or j-to-k (for integer $j \geq 1$ and $k \geq 1$) matching. Fixed-ratio matching can be less efficient than variable-ratio matching because some pruning usually occurs solely to meet this restriction. However, an important goal of matching is simplicity, so the ability to match without having to modify subsequent analysis procedures remains popular. Fixed-ratio matching is also useful in large data sets where observations are plentiful and the primary goal is reducing bias. In practice, most analysts opt for the even more restrictive requirement of one-to-one matching.

In fixed-ratio matching, SATT can be estimated by a simple difference in means between the treated and control groups: $\text{mean}_{i\in\{T=1\}}(Y_i) - \text{mean}_{j\in\{T=0\}}(Y_j)$.

In variable-ratio matching, we can estimate the TE within each matched stratum s by a simple difference in means: $\text{mean}_{i\in s,\{T=1\}}(Y_i) - \text{mean}_{j\in s,\{T=0\}}(Y_j)$.

However, aggregating up to SATT requires weighting, with the stratum-level TE weighted according to the number of treated units. Equivalently, a weighted difference in means can be computed, with weights W such that each treated unit i receives a weight of $W_i = 1$, and each control unit j receives a weight of $W_j = (m_0/m_1)[(m_{s_1})/(m_{s_0})]$, where m_0 and m_1 are, respectively, the number of control and treated units in the data set, and m_{s_1} and m_{s_0} are the number of treated and control units in the stratum containing observation j.[6]

Defining the Number of Units

The size and construction of a matched data set influence the variance of the causal effect estimated from it. Under SATT, the number of treated units remains fixed, and so we measure the data set size by the number of control units. For FSATT, we measure the total number of observations.

For both quantities of interest, we will ultimately use an estimator equal to or a function of the difference in means of Y between the treated and control groups. The variance of this estimator is proportional to $\frac{1}{n_T} + \frac{1}{n_C}$, where n_T and n_C are the number of treated and control units in the matched set. Thus, the variance of the estimator is largely driven by $\min(n_T, n_C)$, and so we will also consult this as an indicator of the size of the data set.

To simplify notation in these different situations, we choose a method of counting from those above and let N denote the number of these units in the original data and n the number in a matched set, with $n \leq N$. In our graphs, we will represent this information as the number of units pruned, which is scaled in the same direction as the variance.

Imbalance Metrics

An *imbalance measure* is a (nondegenerate) indicator of the difference between the multivariate empirical densities of the k-dimensional covariate vectors of treated X_1 and control X_0 units for any data set (i.e., before or after matching). Our concept of a matching frontier, which we define more precisely below, applies to any imbalance measure a researcher may choose. We ease this choice by narrowing down the reasonable possibilities from measures to metrics and then discuss five examples of continuous and discrete families of these metrics. For each example, we give metrics most appropriate for FSATT and SATT when feasible.

Measures versus Metrics. To narrow down the possible measures to study, we restrict ourselves to the more specific concept of an *imbalance metric*, which is a function $d : [(m_0 \times k) \times (m_1 \times k)] \rightarrow [0, \infty]$ with three properties, required for a generic semi-distance:

1. *Nonnegativeness*: $d(X_0, X_1) \geq 0$.
2. *Symmetry*: $d(X_0, X_1) = d(X_1, X_0)$ (i.e., replacing T with $1 - T$ will not affect the calculation).
3. *Triangle inequality*: $d(X_0, X_1) + d(X_1, Z) \geq d(X_0, Z)$, given any k-vector Z.

Imbalance measures that are not metrics have been proposed and are sometimes used, but they add complications such as logical inconsistencies without conferring obvious benefits. Fortunately, numerous imbalance metrics have been proposed or could be constructed.

Adjusting Imbalance Metrics for Relative Importance. Although one can always define a data set that will produce large differences between any two imbalance metrics, in practice the differences among the choice of these metrics are usually not large or at least not the most influential choice in most data analysis problems (Imbens 2004; Zhao 2004). Although we describe continuous and discrete metrics below, more important is the choice of how to scale the variables that go into it. Indeed, every imbalance metric is conditional on a definition of the variables in X, and so researchers should think carefully about what variables may be sufficient in their application.

Analysts should carefully define the measurement of each covariate so that it makes logical sense (e.g., ensuring interval-level measurement for continuous distance metrics) and, most crucially, reflects prior information about its "importance" in terms of its relationship with $Y|T$. In almost all areas of applied research, the most important covariates are well known to researchers—such as age, sex, and education in public health, or partisan identification and ideology in political science. Since bias is a function of both importance and imbalance, and matching is intended to affect the latter (Imai, King, and Stuart 2008), researchers should seek to reduce imbalance more for covariates known to be important. This is easy to do by choosing the right scale for the variables or, equivalently, adjusting weights used in continuous metrics (e.g., the standardization in average Mahalanobis distance or the raw variable scaling in Euclidean distance; see Greevy et al. 2012) or the degree of variable-specific coarsening

[6] See j.mp/CEMweights for further explanation of these weights. Also, for simplicity, we define any reuse of control units to match more than one control as variable-ratio matching.

used in discrete metrics (e.g., H in L_1). We now offer more detail on these continuous and discrete metrics.

Continuous Imbalance Metrics. The core building block of a continuous imbalance metric is a (semi-) distance $D(X_i, X_j)$ between two k-dimensional vectors X_i and X_j, corresponding to observations i and j. For example, the Mahalanobis distance is $\mathrm{D}(X_i, X_j) = \sqrt{(X_i - X_j)S^{-1}(X_i - X_j)}$, where S is the sample covariance matrix of the original data X. The Euclidean distance would result from redefining S as the identity matrix. Numerous other existing definitions of continuous metrics could be used instead. In real applications, scholars should choose variables, the coding for the variables, and the imbalance metric together to reflect their substantive concerns. With this method as with most others, the more substantive knowledge one encodes in the procedure, the better the result will be.

For example, consider data designed to predict the Democratic proportion of the two-party vote in a cross-section of Senate elections ("vote"), with incumbency status as the treatment, controlling for covariates population and the vote in the prior election ("lagged vote"). Clearly, lagged vote will be highly predictive of the current vote, whereas the effect of population will be tiny. Thus, in designing an imbalance metric, we want to be sure to match lagged vote very well, and should only be willing to prune observations based on population when we see gross disparities between treatment and control. For example, if we used Euclidean distance, we could code vote and lagged vote on a scale from 0 to 100 and population on a smaller scale, such as log population. In computing the Euclidean distance between two observations, then, population differences would count as equivalent to relatively small vote differences.

To produce the overall imbalance metric, we must move from the distance between two individual observations to a comparison between two sets of observations, by aggregating binary distance calculations over all the observations. One way to do this is with the *average Mahalanobis imbalance* (AMI) metric, the distance between each unit i and the closest unit in the opposite group, averaged over all units: $D = \mathrm{mean}_i[D(X_i, X_{j(i)})]$, where the closest unit in the opposite group is $X_{j(i)} = \arg\min_{X_j | j \in \{1-T_i\}}[D(X_i, X_j)]$ and $\{1 - T_i\}$ is the set of units in the (treatment or control) group that does not contain i.

Finally, for SATT, it is helpful to have a way to identify the overlap and nonoverlap sets. A natural way to do this for continuous metrics is to define the nonoverlap region as the set of treated units for which no control unit has chosen it as a match. More precisely, denote the (closest) treated unit i that control unit j matches to by $j(i) \equiv \arg\min_{i|i\in\{T=1\}}[D(X_i, X_j)]$. Then define the overlap and nonoverlap sets, respectively, as

$$\mathcal{O} \equiv \{j(i) \mid j \in \{T = 0\}\} \tag{2}$$

and

$$\mathcal{NO} \equiv \{i \mid i \in \{T = 1\} \wedge \{i \notin \mathcal{O}\}\}, \tag{3}$$

where $\wedge$ means "and," connecting two statements required to hold.

Discrete Imbalance Metrics. Discrete imbalance metrics indicate the difference between the multivariate histograms of the treated and control groups, defined by fixed bin sizes H. The results of this calculation depend on the scale of the input covariates, and so the scale can be used to adjust the variables for their likely impact on the outcome variable.

To define these metrics, let $f_{\ell_1 \cdots \ell_k}$ be the relative empirical frequency of treated units in a bin with coordinates on each of the X variables as $\ell_1 \cdots \ell_k$ so that $f_{\ell_1 \cdots \ell_k} = n_{T_{\ell_1 \cdots \ell_k}} / n_T$, where $n_{T_{\ell_1 \cdots \ell_k}}$ is the number of treated units in stratum $\ell_1 \cdots \ell_k$ and n_T is the number of treated units in all strata. We define $g_{\ell_1 \cdots \ell_k}$ similarly among control units. Then, among the many possible metrics built from these components, we consider two:

$$L_1(H) = \frac{1}{2} \sum_{(\ell_1 \cdots \ell_k) \in H} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}| \tag{4}$$

and

$$L_2(H) = \frac{1}{2} \sqrt{\sum_{(\ell_1 \cdots \ell_k) \in H} (f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k})^2}. \tag{5}$$

To remove the dependence on H, Iacus, King, and Porro (2011a) define L_1 as the median value of $L_1(H)$ from all possible bin sizes H in the original unmatched data (approximated by random simulation); we use the same value of H to define L_2. The typically numerous empty cells of each of the multivariate histograms do not affect L_1 and L_2, and so the summation in (4) and (5) each have at most only n nonzero terms.

When used for creating SATT frontiers, these discrete metrics suggest a natural indicator of the nonoverlap region: all observations in bins with either one or more treated units and no controls or one or more control units and no treateds.

With variable-ratio matching, and the corresponding weights allowed in the calculation of L_1, the metric is by definition 0 in the overlap region. With fixed-ratio matching, L_1 will improve as the heights of the treated and control histogram bars within each bin in the overlap region equalize. (In other words, what the weights,

included for variable-ratio matching, do is to equalize the heights of the histograms without pruning observations.)

Constructing Frontiers

Now that we have given our notation and discussed variable- versus fixed-ratio matching, the number of units, and imbalance metrics, we can formally define the frontier and describe algorithms for calculating it.

Definition

Begin by choosing an imbalance metric $d(x_0, x_1)$, a quantity of interest Q (SATT or FSATT), whether to use weights (to allow variable-ratio matching) or no weights (as in fixed-ratio matching) R, and a definition for the number of units U. We will consider all matched data set sizes from the original N, all the way down from $n = N, N-1, N-2 \ldots, 2$.

For quantity of interest SATT, where only control units are pruned, denote $\mathcal{X}_n$ as the set of all $\binom{N}{n}$ possible data sets formed by taking every combination of n rows (observations) from the $(N \times k)$ control group matrix X_0. Then denote the combined set of all sets $\mathcal{X}_n$ as $\mathcal{X} \equiv \{\mathcal{X}_n \mid n \in \{N, N-1, \ldots, 1\}\}$. This combined set $\mathcal{X}$ is (by adding the null set) known as the *power set* of rows of X_0, containing (a gargantuan) 2^N elements. For example, if the original data set contains merely $N = 300$ observations, the number of elements of this set exceeds current estimates of the number of elementary particles in the universe. The task of finding the frontier requires identifying a particular optimum over the entire power set. Using a brute force approach of trying them all and choosing the best is obviously infeasible even with all the world's computers turned to the task. A key contribution of this article is a set of new algorithms that make possible finding this exact same frontier in only a few minutes on an ordinary personal computer.

To be more specific, first identify an element (i.e., data set) of $\mathcal{X}_n$ with the lowest imbalance for a given matched sample size n, and the choices of Q, U, and R:

$$x_n = \underset{x_0 \in \mathcal{X}_n}{\arg\min} \; d(x_0, x_1), \text{ given } Q, R, \text{ and } U \qquad (6)$$

where for convenience when necessary we define the arg min function in the case of nonunique minima as a random draw from the set of data sets with the same minimum imbalance. We then create a set of all these minima $\{x_N, x_{N-1}, \ldots, 1\}$, and finally define the *matching frontier* as the subset $\mathcal{F}$ of these minima after imposing monotonicity, which involves eliminating any element that has higher imbalance with fewer observations:

$$\mathcal{F} \equiv \{x_n \mid (n \in \{N, N-1, \ldots, 1\}) \wedge (d_{n-1} \leq d_n)\}, \qquad (7)$$

where $d_n = d(x_n, x_1)$. We represent a frontier by plotting the number of observations pruned $N - n$ horizontally and d_n vertically.

For simplicity, we will focus on SATT here, but our description also generalizes to FSATT by defining $\mathcal{X}_n$ as the set of all combinations of the entire data matrix $(X_0', X_1')'$ taken n at a time.

Algorithms

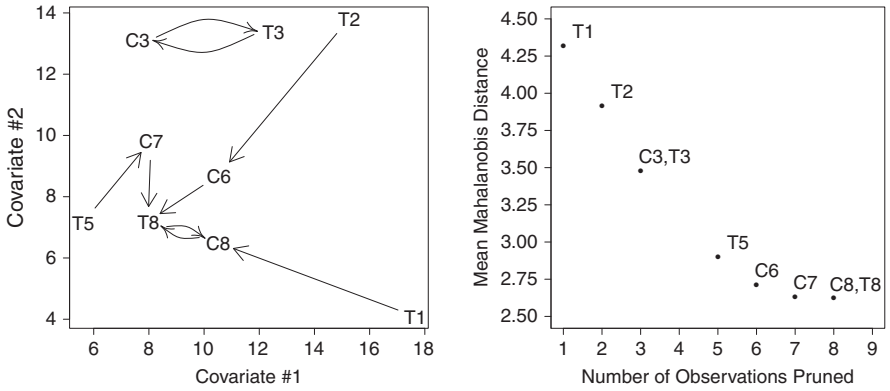
Calculating the frontier requires finding a data subset of size n with the lowest imbalance possible, chosen from the original data of size N for each possible n ($N > n$)—given choices of the quantity of interest (and thus the definition of the units to be pruned), fixed- or variable-ratio matching, and an imbalance metric, along with the monotonicity restriction.

Adapting existing approaches to algorithms for this task is impractical. The most straightforward would involve directly evaluating the imbalance of the power set of all possible subsets of observations. For even moderate data set sizes, this would take considerably longer than the expected lives of most researchers (and for reasonably sized data sets, far longer than life has been on the planet). Another approach could involve evaluating only a sample of all possible subsets, but this would be biased, usually not reaching the optimum (for the same reason that a maximum in a random sample is an underestimate of the maximum in the population). Finally, adapting general purpose numerical optimization procedures designed for similar but different purposes would take many years, and they are not guaranteed to reach the optimum.[7]

The contribution of our algorithms, then, is not that they can find the optimum, but that they can find it fast. Our solution is to offer analytical rather than numerical solutions to this optimization problem by leveraging the properties of specific imbalance metrics.

[7]We are indebted to Diamond and Sekhon (2012) for their idea of "genetic matching," which seeks to approximate one point on the frontier (i.e., one matched sample size). For this point, imbalance will usually be higher and never lower than the frontier, and so the approach proposed here dominates. We also find, using data from Sekhon (2011) as an example, that computing the entire (optimal) frontier with our approach is about 100 times faster than it takes genetic matching to compute one point (approximating, but not guaranteed to be on, the frontier). For each individual point, our approach is about 1,000,000 times faster.

FIGURE 1 A Demonstration of the Continuous Metric FSATT Algorithm



Note: The left panel gives an example data set with two covariates, one on each axis. The numbers indicate the order in which observations are pruned, the arrows denote the observation in the opposite treatment regime to which each unit is matched, and "T" or "C" denote treatment and control, respectively. The right panel displays the frontier for this data set, where points are labeled to correspond with the left plot.

We now outline algorithms we developed for calculating each of four families of matching frontiers, with many possible members of each. We leave to future research the derivation of algorithms for other families of matching frontiers (defined by choosing among permutations of the choices defined in the previous subsection, and finding a feasible algorithm). In all cases with SATT frontiers, we first remove the nonoverlap set and then compute the remaining frontier within the overlap set. FSATT frontiers do not require this separate step.

Continuous, FSATT, Variable Ratio. Our first family of algorithms is based on the average continuous imbalance metric for FSATT with variable-ratio matching. We illustrate this with the AMI metric, although any other continuous metric could be substituted. Define N as the number of units in the original data, n as the number that have not yet been pruned, and D_n as the current matched data set. Then our algorithm requires five steps:

1. Begin with $N = n$.
2. Match each observation i ($i = 1, \ldots, N$) to the nearest observation $j(i)$ in the opposite treatment condition $\{1 - T_i\}$; this matched observation has index $j(i) = \arg\min_{j \in \{1-T_i\}}[D(X_i, X_j)]$, and distance $d_{j(i)} \equiv D(X_i, X_{j(i)})$.
3. Calculate AMI for D_n.
4. Prune from D_n the unit or units with distance $d_{j(i)}$ equal to $\max(d_{j(i)} | i \in D_n)$. Redefine n as the number of units remaining in newly pruned data set D_n.
5. If $n > 2$ and AMI > 0, go to Step 3; else stop.

This is a greedy algorithm, but we now prove that it is optimal by proving two claims that, if true, are sufficient to establish optimality. Below, we offer a simple example.

Claim 1. *For each unit* i *remaining in the data set, the nearest unit (or units) in the opposite treatment regime* j(i) *(defined in Step 2) remains unchanged as units are pruned in Step 4.*

Claim 2. *The subset D_n with the smallest AMI, among all $\binom{N}{n}$ subsets of D_N, is that defined in Step 4 of the algorithm.*

Given Claim 1, Claim 2 is true by definition since D_n includes the n smallest distances within D_N.

We prove Claim 1 by contradiction. First assume that the next unit to be pruned is the nearest match for a unit not yet pruned (hence requiring the data set to be rematched). However, this is impossible because the unit not yet pruned would have had to have a distance to the pruned unit lower than the maximum, in which case it would have been matched to that unit originally, which is not the case. This proves Claim 1. By induction, we then have a proof of optimality.

To illustrate, consider the simple data set in the left panel of Figure 1, with five treated units (denoted by the letter "T") and four control units (denoted by the letter "C"), each measured on two covariates (one on each axis). We denote, with an arrow coming from each unit, the closest unit (measured in Mahalanobis distance) in the opposite treatment regime.

The algorithm is illustrated by removing the observations in numerical order in the left panel of Figure 1, starting with observation T1. The right panel gives the

frontier for these data, numbered with the observation to be removed next.

The figure also illustrates the special case of "mutual minima," which is a treated unit that has as its nearest match a control unit that, in turn, has as its closest match the same treated unit. In this situation, we remove both (or more if they are also tied) units at the same time, as per Step 4. This is illustrated in the left panel of Figure 1 by two arrows between the pair of observations marked 3 (and also 8, when only two units remain). Technically, this means that some points on the frontier are not represented, such as if we wished to prune exactly three observations in this simple example. Although we could fill in these missing points by enumerating and checking all possible data sets this size, we find that omitting them, and thus saving substantial computational time, is almost costless from a practical point of view: A researcher who wants a data set of a particular sample size would almost always be satisfied with a very slightly different sample size. With a realistically sized data set, the missing points on the frontier, like discrete points in a (near) continuous space, are graphically invisible and for all practical purposes substantively irrelevant.[8]

Put differently, the intuition as to why the greedy algorithm is also optimal is that at every point along the frontier, the closest match for each observation in the remaining set is the same as that in the full set, which implies that (1) each observation contributes a fixed distance to the average distance for the entire portion of the frontier in which it remains in the sample and (2) observations do not need to be rematched as others are pruned.

Continuous, SATT, Variable Ratio. The SATT frontier is identical to the FSATT frontier when the SATT requirement of keeping all treated units is not binding. This occurs when the $\max_i d_{j(i)}$ calculation in the FSATT algorithm leads us to prune only control units. Usually, this is the case for some portion of the frontier, but not the whole frontier. When the calculable part of this SATT frontier is not sufficient, part of each matched data set along the rest of the frontier will have a nonoverlap region, requiring extrapolation. In these situations, we recommend using the FSATT algorithm for the overlap region, extrapolating to the remaining nonoverlap region and combining the results as per Equation (1).

[8]It is important to note that the missing points on the frontier cannot violate the monotonicity constraint. If the point after the gap contains n observations, those n observations are the n observations with the closest matches. Therefore, there can be no set of $n+1$ observations (say, in the gap) with a lower mean distance than that in the n observations contained at the defined point because the additional observation must have a distance equal to or greater than the greatest distance in the existing n observations.

Discrete, FSATT, Variable Ratio. As a third family of frontiers, we can easily construct a discrete algorithm for FSATT with variable-ratio matching, using L_1 for exposition. With variable-ratio matching in discrete metrics, weighting eliminates all imbalance in the overlap set, making this frontier a simple step function with only two steps. The first step is defined by L_1 for the original data, and the second is L_1 after the nonoverlap region has been removed. Within each step, we could remove observations randomly one at a time, but since imbalance does not decline as a result, it makes more sense to define the frontier for only these two points on the horizontal axis.

If the binning H is chosen to be the same as the coarsening in coarsened exact matching (CEM), the second step corresponds exactly to the observations retained by CEM (Iacus, King, and Porro 2011a, 2011b).

Discrete, SATT, Fixed Ratio. A final family of frontiers is for discrete metrics, such as L_1 or L_2, for quantity of interest SATT with fixed-ratio matching. To define the algorithm, first let b_{iT} and b_{iC} be the numbers, and $p_{iT} = b_{iT}/n_T$ and $p_{iC} = b_{iC}/n_C$ be the proportions, of treated and control units in bin i ($i \in \{1, \ldots, B\}$) in the L_1 multivariate histogram, where $n_T = \sum_{i=1}^{B} b_{iT}$ and $n_C = \sum_{i=1}^{B} b_{iC}$ are the total numbers of treated and control units.

To prune k observations optimally (i.e., with minimum L_1 imbalance) from a data set of size N, we offer this algorithm:

1. Define $p'_{iC} = b_{iC}/(n_C - k)$.
2. Prune up to k units from any bin i where after pruning $p'_{iC} \geq p_{iT}$ holds.
3. If k units have not been pruned in Step 2, prune the remaining k' units from the bins with the k' largest differences $p_{iC} - p'_{iT}$.

An optimal frontier can then be formed by applying this algorithm with $k = 1$ pruned and increasing until small numbers of observations result in nonmonotonicities in L_1.

The discreteness of the L_1 imbalance metric means that multiple data sets have equivalent values of the imbalance metric for each number of units pruned. Indeed, it is possible with this general algorithm to generate one data set with $k-1$ pruned and another with k pruned that differ with respect to many more than one unit. In fact, even more complicated is that units can be added, removed, and added back for different adjacent points on the frontier. This, of course, does not invalidate the algorithm, but it would make the resulting frontier difficult to use in applied research. Thus, to make this approach

easier to use, we choose a greedy algorithm that is a special case of this general optimal algorithm. The greedy algorithm is faster, but more importantly, it by definition never needs to put a unit back in a data set once it has been pruned.

Our greedy algorithm, which we show to be optimal, is as follows. Starting with the full data set $N = n$,

1. Compute and record the value of L_1 and the number of units n.
2. Prune a control unit from the bin with the maximum difference between the proportions of treated and control units, such that there are more controls than treateds. That is, prune a unit from bin $f(i)$, where
$$f(i) = \underset{i \in \{n_C > n_T\}}{\arg\max} |p_{ic} - p_{it}|. \quad (8)$$
3. If L_1 is larger than the previous iteration, stop; otherwise, go to Step 1.

To understand this algorithm intuitively, first note that deleting a control unit from any bin with more controls than treateds changes L_1 by an equal amount (because we are summing over bins normalized by the total number of controls, rather than by the number of controls in any particular bin). When we delete a control unit in one bin, the relative size of all the other bins increases slightly because all the bins must always sum to 1. Deleting controls from bins with the greatest relative difference, as we do, prevents the relative number of treated units from ever overtaking the relative number of controls in any bin and guarantees that this greedy algorithm is optimal.

To illustrate the greedy version of this optimal algorithm, Figure 2 gives a simple univariate example. Panel 0 in the top left represents the original data set with a histogram in light gray for controls and dark gray for treated units. The L_1 imbalance metric for Panel 0 is reported in the frontier (Figure 2, bottom right) marked as "0." The unit marked in black in Panel 0 is the next control unit to be removed, in this case because it is in a bin without any treated units.

Then Panel 1 (i.e., where one observation has been pruned) removes the black unit from Panel 0 and renormalizes the height of all the bars with at least some control units so that they still sum to 1. As indicated by the extra horizontal lines, reflecting the heights of control histogram bars from the previous panel, the heights of each of the remaining light gray control histogram bars have increased slightly. The "1" point in the bottom-right panel in Figure 2 plots this point on the frontier. The black piece of Bin 5 in Panel 1 refers to the next observation to be removed, in this case because this bin has the largest difference between control and treateds, among those with more controls than treateds (i.e., Equation 8). Panels 2–4 repeat the same process as in Panel 1, until we get to Panel 4, where no additional progress can be made and the frontier is complete.

Applications

We now demonstrate our approach in two substantive applications. The first uses a SATT frontier, so a direct comparison can be made between experimental and observational data sets. The second uses an FSATT frontier in a purely observational study, and thus we focus on understanding the new causal quantity being estimated.

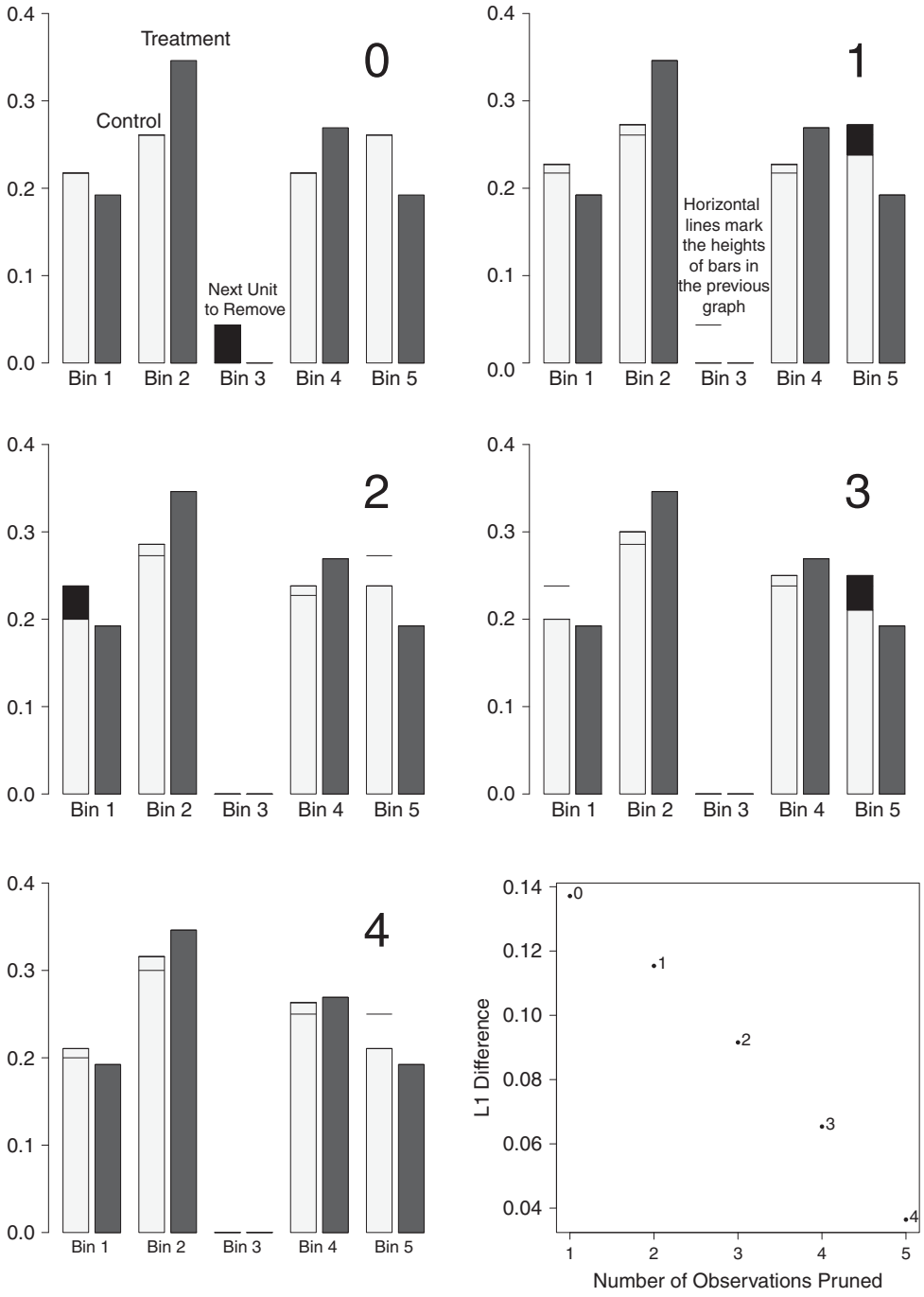
Job Training

We begin with a data set compiled from the National Supported Work Demonstration (NSWD) and the Current Population Survey (Dehejia and Wahba 2002; Lalonde 1986). The first was an experimental intervention, whereas the second is a large observational data collection appended to the 185 experimental treated units and used in the methodological literature as a mock control group to test a method's ability to recover the experimental effect from observational data. Although our purpose is only to illustrate the use of the matching frontier, we use the data in the same way.

As is standard in the use of these data, we match on age, education, race (black or Hispanic), marital status, whether or not the subject has a college degree, earnings in 1974, earnings in 1975, and indicators for whether or not the subject was unemployed in 1974 and 1975. Earnings in 1978 is the outcome variable. To ensure a direct comparison, we estimate the SATT fixed-ratio frontier, and thus prune only control units. We give the full matching frontier in the top-left panel of Figure 3, and, in the two lower panels, the effect estimates for every point on the frontier (the lower-left panel displays estimates over the full frontier, and the lower-right panel zooms in on the final portion of the frontier). There is no risk of selection bias in choosing a point on the frontier based on the estimates, so long as the entire set is presented, as we do here.

These data include 185 treated (experiment) units and 16,252 control units, so pruning even the vast majority of the control units will not reduce the variance much and could help reduce imbalance (see the section "Defining the Number of Units"). In the present data, this insight is especially important given that the

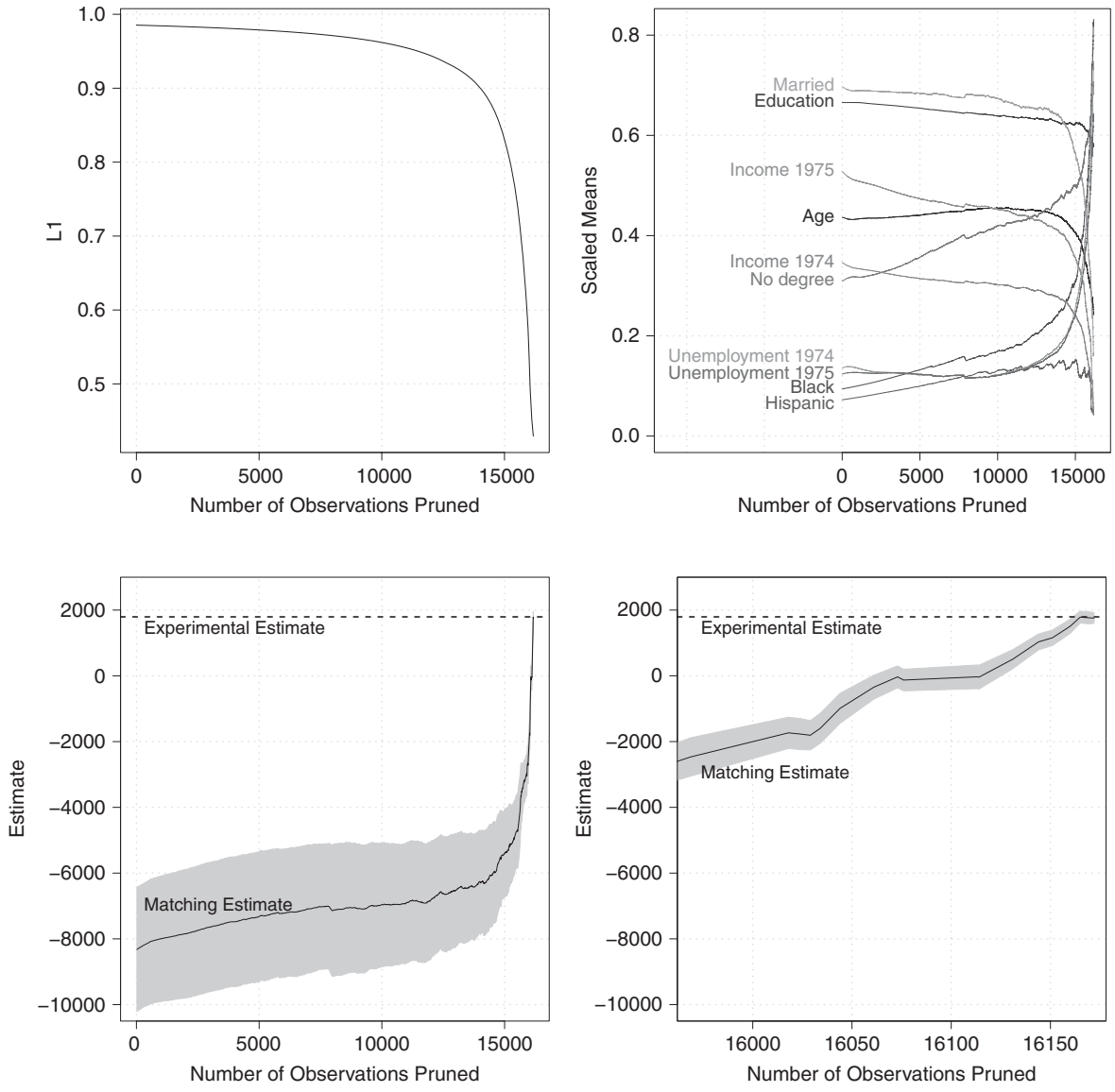
FIGURE 2 A Demonstration of the L_1 SATT Algorithm



largest trade-off between pruning and imbalance occurs after most of the observations are pruned; this can be seen at the right of the left panel of Figure 3, where the frontier drops precipitously.

The bottom left panel gives the causal effect estimates, where the apparent advantages of pruning most of the cases can be seen. Over most of the range of the frontier, causal estimates from the (badly matched) data do not move much from the original unmatched estimate, −$8,334, which would indicate that this job training program was a disaster. Then after the fast drop in imbalance (after pruning about 15,000 mostly irrelevant observations), the estimates rise fast and ultimately intersect with the experimental estimate of the training program, producing a benefit of $1,794 per trainee (denoted by the red horizontal line). Overall, the frontier reveals the

FIGURE 3 The Effect of a Job Training Program on Wages



Note: The top-left panel is the L_1 frontier for the job training data, beginning at the left before pruning with $n = 16,252$ (nonexperimental) control units. The top-right panel displays covariate means across the frontier. The bottom-left panel displays causal effect estimates along that frontier, with Athey-Imbens model dependence intervals calculated at each point. The bottom-right panel displays the final points on the frontier, effectively "zooming" in on the portion where imbalance improves rapidly. The horizontal dashed line is the estimated treatment effect from the experimental sample.

whole range of possible conclusions as a function of the full bias-variance trade-off. The bottom panels also gives Athey-Imbens model dependence intervals (Athey and Imbens 2015) around the point estimates;[9] the widths of these are controlled by the model dependence remaining in the data and so decrease as balance improves across the frontier. Correspondingly, the largest change in model dependence occurs near the end of the frontier, where imbalance improves the most.

[9]Athey and Imbens (2015) propose "a scalar measure of the sensitivity of the estimates to a range of alternative models." To compute this measure, investigators estimate the quantity of interest with a base model, after which the quantity of interest is estimated in subsamples divided according to covariate values. The deviation of these subsample estimates from the base estimate is then a measure of model dependence.

Sex and Judging

For our second example, we replicate Boyd, Epstein, and Martin (2010), who offer a rigorous analysis of the

effect of sex on judicial decision making. They first review the large number of theoretical and empirical articles addressing this question and write that "roughly one-third purport to demonstrate clear panel or individual effects, a third report mixed results, and the final third find no sex-based differences whatsoever." These prior articles all use similar parametric regression models (usually logit or probit) and related data sets. To tame this remarkable level of model dependence, they introduce nearest neighbor propensity score matching to this literature, finding no effects of sex on judging in every one of 13 policy areas except for sex discrimination, which makes good sense substantively. The authors also offer a spirited argument for bringing "promising developments in the statistical sciences" to important substantive questions in judicial politics, and so we follow their lead here, too. We thus follow their inclinations but with methods developed after the publication of their article, by seeing whether our more powerful approach can detect results not previously possible.

Boyd, Epstein, and Martin (2010) motivate their study by clarifying four different mechanisms through which sex might influence judicial outcomes, each with distinct empirical implications. First, *different voice* implies that "males and females develop distinct worldviews and see themselves as differentially connected to society." This account suggests that males and females should rule differently across a broad range of issues, even those with no clear connection to sex. Second, the *representational* mechanism posits that "female judges serve as representatives of their class and so work toward its protection in litigation of direct interest." This theory predicts that males and females judge differently on issues of immediate concern to women. Third, the *informational* account argues that "women possess unique and valuable information emanating from shared professional experiences." Here, women judge differently on the basis of their unique information and experience, and so they might differ from men on issues over which they have distinct experiences, even if not related to sex. Finally, the *organizational* theory posits that "male and female judges undergo identical professional training, obtain their jobs through the same procedures, and confront similar constraints once on the bench." This theory predicts than men and women do not judge differentially.

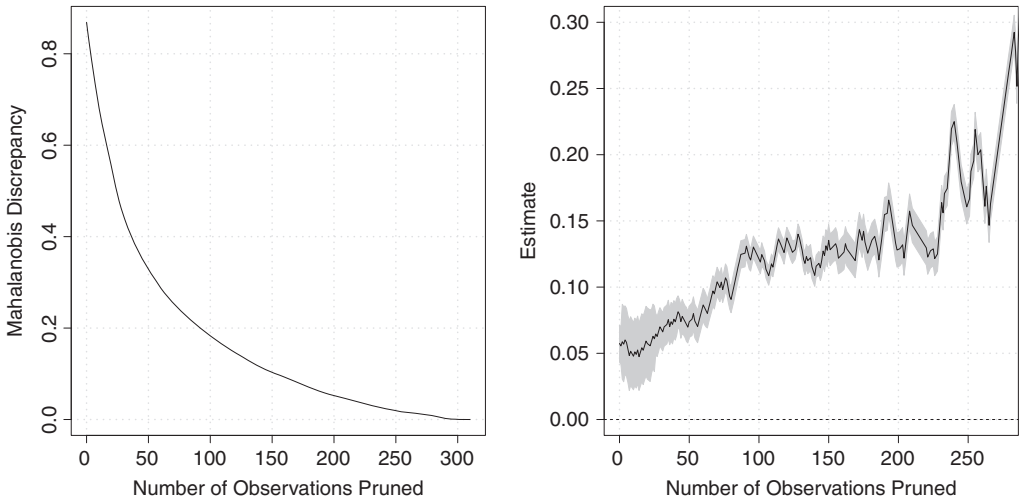
Boyd, Epstein, and Martin (2010) argue that their results—that males and females judge differently *only* in sex discrimination cases—are "consistent with an information account of gendered judging." Of course, their results are also consistent with representational theories. Indeed, as Boyd, Epstein, and Martin (2010) argue, women might judge differently on sex discrimination because they have different information as a result of shared experiences with discrimination. But it is also possible that women judge differently on sex discrimination as a way to protect other women, consistent with representational accounts.

One way to use our new methodological approach is to attempt to distinguish between these conflicting interpretations of the effects of sex. To do so, we analyze cases on an issue for which we expect to observe a difference if and only if the *informational* account is true—that is, an issue area where the unique experiences of women might lead to informational differences between men and women but that nonetheless does not directly concern the interests of women. For this analysis, we consider cases related to discrimination on the basis of race. Because women have shared experiences with discrimination, they have informational differences from men relevant to this issue area. However, judgments on racial discrimination do not have direct consequences for women more broadly. This issue area is the only such issue area that allows us to distinguish between these two accounts and for which we also have a suitable amount of available data.

Thus, using their data, we reanalyze race discrimination cases made on the basis of Title VII. In their original analysis, Boyd, Epstein, and Martin (2010) found a null effect with this issue area, both before and after matching. We now show, with our method, which enables us to analyze data with less dependence and bias than previous matching approaches, that female judges rule differently on race discrimination cases. We show that differences in male and female judgments are at least in part due to informational differences.

We arranged the data from Boyd, Epstein, and Martin (2010) so that the unit of analysis is the appellate court case, the level at which decisions are made. For example, the fourth observation in our data set is *Swinton v. Potomac Corporation*, which was decided in 2001 by the Ninth Circuit Court of Appeals, at the time composed of Judges William A. Fletcher, Johnnie B. Rawlinson, and Margaret M. McKeown. Our treatment is whether or not at least one female judge was included in the three-judge panel. In *Swinton v. Potomac Corporation*, Judges Rawlinson and McKeown are female, and so this observation is in the treatment group. For each appellate court case, we use the following covariates: (1) median ideology as measured by Judicial Common Space scores (Epstein et al. 2007; Giles, Hettinger, and Peppers 2001), (2) median age, (3) an indicator for at least one racial minority among the three judges, (4) an indicator for ideological direction of the lower court's decision, (5) an indicator for whether a majority of the judges on the three-judge panel were nominated by Republicans, and (6) an indicator for whether a

FIGURE 4 The Effect of Sex on Judges in Title VII Race Discrimination Cases



(a) Title VII Race Imbalance-*n* Frontier **(b)** Title VII Race Causal Effects Estimates

Note: Panel (a) displays the imbalance-*n* frontier, and Panel (b) shows estimates of the causal effect across that same frontier. The shaded region is the Athey-Imbens model dependence interval.

majority of the judges on the panel had judicial experience prior to their nomination. In *Swinton v. Potomac Corporation*, for example, the median ideology was at the 20th percentile in the distribution of judges who ruled on a Title VII race discrimination case and in the 10th percentile in the distribution of median ideologies, where lower scores are more liberal and higher scores are more conservative. This is unsurprising, as all three judges were nominated by President Clinton in either 1998 or 2000. Our outcome is the ideological direction of the decision—either liberal or conservative—and unsurprisingly, the ruling on *Swinton v. Potomac Corporation* was a liberal one. For our analysis, we use these six covariates to construct a Mahalanobis frontier for the estimation of FSATT.

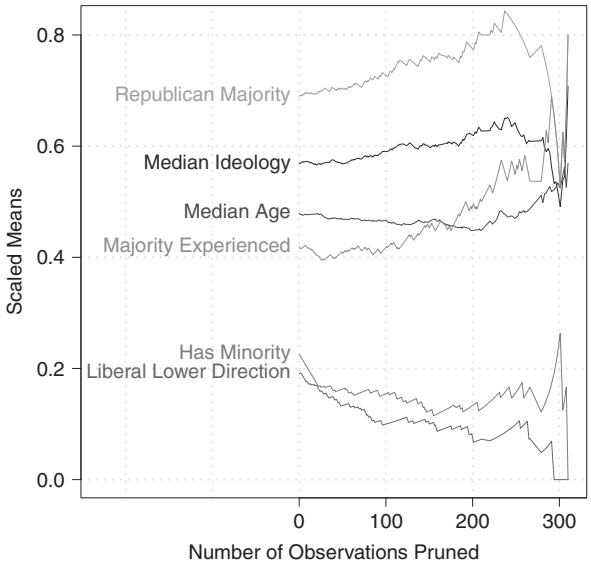
We present the matching frontier in Figure 4, Panel (a); in contrast to our previous example, most of the reduction in imbalance for observations pruned happens early on, but substantial imbalance reduction continues through the entire range. Our substantive results can be found in Panel (b), which indicates that having a female judge increases the probability of a liberal decision, over the entire range. The vertical axis quantifies the substantial effect we see in terms of the reduction in probability, from about 0.05 with few observations pruned and higher levels of model dependence to 0.25 with many pruned and lower levels. Most importantly in this case, the point estimate of the causal effect increases as balance improves and we zero in on a data subset more closely approximating a randomized experiment. Correspondingly, model dependence, indicated by the shaded interval, decreases.

Because we pruned both treated and control units in this example, we must carefully consider how the quantity of interest changes as balance improves. Interestingly, the best balance exists within more conservative courts (and thus the effect we find in Figure 4 is among these courts). To see this, note that in Figure 5, which plots the means of each covariate as observations are pruned, variables associated with the court ideology changed the most, and all moved in a conservative direction.

More specifically, *Republican majority* has the largest difference in means, followed by *median ideology* and *median age*. That makes sense, as we expect all-male panels (those assigned to control) to be more conservative on average, if only because Democratic presidents are more likely to appoint women than Republicans. In our data, 7% of judges appointed by Republicans are women, compared with 25% of judges appointed by Democrats. The key issue is that liberal courts with no female judges are rare, and so the data do not admit reasonable inferences among this subset of data (e.g., 31% of courts with a Republican majority have at least one woman compared with 50% of courts without a Republican majority).

Our matching technique thus successfully identifies a subset of data with balance, and this subset is where women are assigned to conservative courts as well as a suitable comparison group of conservative courts without a woman. The causal effect of sex on judging in conservative courts is not the only question of interest in the literature, but it is certainly one question of interest, and, as it happens, it is one in these data that offers a

FIGURE 5 Covariate Means Across the Title VII Race Discrimination Frontier



(a) Title VII Race Covariate Means

Note: The figure displays the means of each covariate as observations are pruned and balance improves. Note the large change in variables measuring the ideology of the court.

reasonably secure answer. Thus, for example, after pruning 210 appellate court cases, the difference in the median Judicial Common Space score for treated and control units is 0.02, compared to 0.1 in the full data. However, interestingly, the average median score for controls is essentially constant as we prune observations, changing from 0.123 to 0.163. In contrast, the average median score for treated units (courts with at least one woman) goes from 0.02 to 0.143, which is to say it rises to the level of conservatism displayed by courts assigned to the control group.

These results suggest an especially interesting conclusion to supplement the original findings in Boyd, Epstein, and Martin (2010), namely, that the mechanism for differences in judicial decision making is at least in part informational, perhaps in addition to being representational, even among the most conservative courts.

Concluding Remarks

Matching methods provide a simple yet powerful way to control for measured covariates. However, in practice, implementing matching methods can be frustrating because they require analysts to jointly optimize balance and sample size through a complicated, poorly guided process of tweaking matching algorithms to obtain better matched samples. The matching frontier we describe here offers the first simultaneous optimization of both balance and sample size while retaining much of the simplicity that made matching attractive in the first place.

With the approach offered here, once a researcher chooses an imbalance metric and set of covariates, all analysis is automatic. This is in clear distinction to the best practices recommendations for prior matching methods. However, although the choice of a particular imbalance metric does not usually matter that much, the methods offered here do not free one from the still crucial task of choosing the right set of pretreatment control variables, coding them appropriately so that they measure what is necessary to achieve ignorability or, in the case of FSATT, from understanding and conveying clearly to readers the quantity being estimated.

Our approach generalizes directly for multivalued treatments without major changes in any of the technology described. A larger change, and potentially fruitful topic for future research, would be combining our approach to the frontier with formal sensitivity testing for causal inference.

References

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

Austin, Peter C. 2008. "A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003." *Journal of the American Statistical Association* 27(12): 2037–49.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies*, ed. E. Stromsdorfer and G. Farkas. Vol. 5. San Francisco: Sage.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. "Untangling the Causal Effects of Sex on Judging." *American Journal of Political Science* 54(2): 389–411.

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

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96(1): 187–99.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84(1): 151–61.

Diamond, Alexis, and Jasjeet 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): 932–45.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal, and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law, Economics, and Organization* 23(2): 303–25.

Giles, Micheal W., Virginia A. Hettinger, and Todd Peppers. 2001. "Picking Federal Judges: A Note on Policy and Partisan Selection Agendas." *Political Research Quarterly* 54(3): 623–41.

Greevy, Robert A., Carlos G. Grijalva, Christianne L. Roumie, Cole Beck, Adriana M. Hung, Harvey J. Murff, Xulei Liu, and Marie R. Griffin. 2012. "Reweighted Mahalanobis Distance Matching for Cluster-Randomized Trials with Missing Data." *Pharmacoepidemiology and Drug Safety* 21(S2): 148–54.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* Evaluation of Training and Other Social Programmes (Oct., 1997), 64(4): 605–654.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011a. "Multivariate Matching Methods That Are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106(493): 345–61. http://gking.harvard.edu/files/abs/cem-math-abs.shtml.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011b. "Replication Data for Causal Inference without Balance Checking: Coarsened Exact Matching." http://hdl.handle.net/1902.1/15601. Murray Research Archive [Distributor], V1 [Version].

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171(2): 481–502. http://gking.harvard.edu/files/abs/matchse-abs.shtml.

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

Imbens, Guido W., and Donald B. Rubin. 2009. "Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction." 2015. New York: Cambridge University Press.

King, Gary, Christopher Lucas, and Richard A. Nielsen. 2016. *MatchingFrontier: Automated Matching for Causal Inference.* R package version 2.0.0.

King, Gary, Lucas Christopher, Richard A. Nielsen 2016. "Replication Data for: The Balance-Sample Size Frontier in Matching Methods for Causal Inference", doi:10.7910/DVN/SURSEO, Harvard Dataverse, V1 [UNF:6:N+6bvznjSZ2ZDl/OMFDocQ==]

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76(4): 604–20.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

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

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387): 515–24.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3): 808–40.

Rubin, Donald B. 2010. "On the Limitations of Comparative Effectiveness Research." *Statistics in Medicine* 29(19): 1991–95.

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal of Statistical Software* 42.

Stuart, Elizabeth A. 2008. "Developing Practical Recommendations for the Use of Propensity Scores: Discussion of 'A Critical Appraisal of Propensity Score Matching in the Medical Literature between 1996 and 2003'." *Statistics in Medicine* 27(12): 2062–65.

VanderWeele, Tyler J., and Miguel A. Hernan. 2012. "Causal Inference under Multiple Versions of Treatment." *Journal of Causal Inference* 1: 1–20.

VanderWeele, Tyler J., and Ilya Shpitser. 2011. "A New Criterion for Confounder Selection." *Biometrics* 67(4): 1406–13.

Zhao, Zhong. 2004. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence." *Review of Economics and Statistics* 86(1): 91–107.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix