Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects

Chad Hazlett – University of California Los Angeles

Abstract

Methods such as matching and weighting for causal effect estimation attempt to make to achieve multivariate balance, making the distribution of covariates for untreated units the same as that for the treated units. However, they often cannot fully achieve this, and estimates of the average treatment effect on the treated (ATT) are biased when any function of the covariates that influences the non-treatment potential outcome has a different mean for the treated and untreated units. Kernel balancing targets a simpler requirement for unbiased ATT estimation: that the mean non-treatment potential outcome for the treated and control groups are equal in expectation. This can be achieved under mild smoothness assumptions on the regression surface for the non-treatment potential outcome. Despite this different goal, these weights produced by kernel balancing are nevertheless interpretable as (1) those that ensure a particular approximation of the multivariate distribution of the covariates is the same for the treated and controls, and (2) a form of stabilized inverse propensity score weights that does not require a model of the treatment. An `R` package, `KBAL` provides an implementation of this approach.

Chad Hazlett, Assistant Professor of Statistics and Political Science. E-mail: chazlett@ucla.edu.
Website: www.chadhazlett.com

1. Introduction

Kernel balancing uses a kernel to construct a higher dimensional transformation of the original data, then chooses weights to achieve equal means for the treated and control groups on this transformed version of the data. While simple, this method makes several contributions to existing methodology.

First, while matching, covariate balancing weights, and propensity score methods can be understood as seeking to make the multivariate distribution of the covariates for the controls identical to that of the treated units, this is more than is required for unbiased estimation of the average treatment effect on the treated (ATT). The goal of kernel balancing is, first, to ensure that the non-treatment potential outcome has the same mean for the treated and control group under the most general conditions possible. This achieves the minimal requirement for unbiased estimation of the ATT for a simple weighted difference in means estimator.

Second, in the matching and covariate-balancing literatures, there is no clear answer to the question of "on what functions of the covariates should the investigator check balance?" Yet, failure to obtain balance on some function of the covariates (such smooth but non-linear joint functions of two or more covariates) will cause biased ATT estimates if that function of the covariates also predicts the outcome (illustrated in 2 below). Kernel balancing ensure the treated and control groups will have the same means not only on the covariates, but on a wide range of non-linear functions of the covariates. In so doing, it provides a principled answer to the question of what functions of the covariates to balance.

Third, kernel balancing finds weights such that multivariate densities of the treated and control are equal, as evaluated at every observation in the dataset, for a particular choice of kernel smoother used to estimate those densities. The weights are also equivalent to a non-parametric form of stabilized inverse propensity score weights. Thus, while focusing first on the minimum requirement for unbiased ATT estimation, the method also achieves the goals for which matching, weighting, and propensity score have traditionally been employed.

In what follows, section 2 first provides an illustration of the risk of bias under existing methods and the benefit of the proposed solution. Section 3 describes the basic idea behind kernel balancing. The discussion in section 4 places this approach in the context of other matching, weighting, propensity score methods, provides further beneficial properties of the method, and describes additional implementation details. Section 5 offers an empirical application, with further applications and other materials available in the appendix.

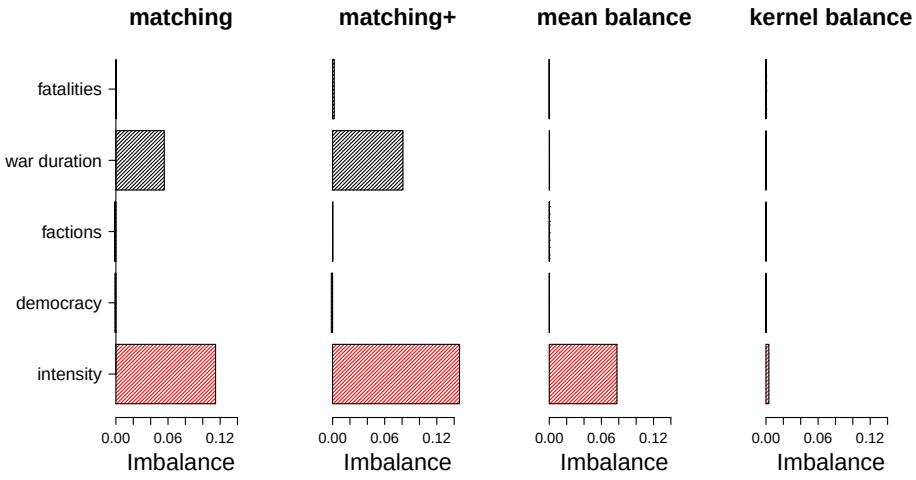
2. Motivating Example

To illustrate the risk of bias under existing methods, I begin with a simple motivating example. This example uses simulated data, but in the context of a real world estimation challenge. Suppose we are interested in the question of whether peacekeeping missions deployed after civil wars are effective in lengthening the duration of peace (*peace years*) after the war's conclusion (e.g. Fortna 2004; Doyle and Sambanis 2000). However, the "treatment" – peacekeeping missions (*peacekeeping*) – is not randomly assigned. Rather, missions are more likely to be deployed in certain situations, which may differ systematically in their expected *peace years* even in the absence of a peacekeeping mission. To deal with this, we collect four pre-treatment covariates that describe each case: the duration of the preceding war (*war duration*), the number of fatalities ($fatalities$), democracy level prior to the peacekeeping mission (*democracy*), and a measure of the number of factions or sides in the civil war ($factionalism$). We are interested in the average treatment effect on the treated (ATT), which

is the mean number of *peace years* experienced by countries that received *peacekeeping*, minus the average number of *peace years* for this group had they not received peacekeeping missions.

Further, let us suppose that there are no unobserved confounders, but that peacekeeping missions are deployed on the basis of a conflict's *intensity*, which equals $\frac{fatalities}{war\,duration}$. In particular, missions are more likely to be deployed where conflicts were higher in intensity. Suppose the outcome of interest, *peace years*, is also a function of *intensity*, with more intense conflicts leading to longer average *peace years*. This is reasonable if, for example, more intense wars indicate greater dominance by one side, leading to a lower likelihood of resurgence in each subsequent year. In this example, *peace years* is only a function of *intensity*, and not of *peacekeeping*, implying a true treatment effect of zero.

Figure 1: Imbalance on a function of the covariates

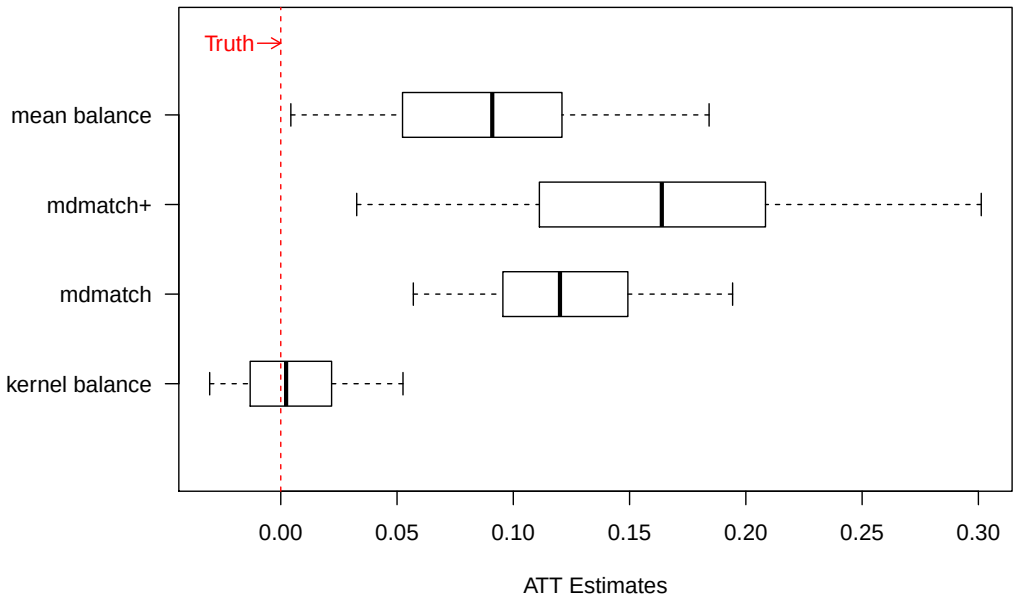


Mean imbalances on included covariates and $intensity = \frac{fatalities}{war\,duration}$, which determines both assignment of the treatment (*peacekeeping*) and the outcome (*peaceyears*). *Matching:* Mahalanobis distance matching on the original covariates alone leaves a substantial imbalance on *war duration*. More problematically, it shows a large imbalance on *intensity*. *Matching+:* Mahalanobis distance matching with squared terms and all pairwise multiplicative worsens imbalance, particularly on *intensity*. *Mean balance*: Entropy balancing on the original covariates achieves essentially perfect mean balance on these, but only a small improvement in balance on *intensity*. *Kernel balance* obtains mean balance on a wide range of smooth functions of the included covariates, obtaining balance *intensity* despite not including it in the algorithm.

How well do existing techniques achieve equal means for the treated and controls ("mean balance"), both on the original four covariates and on *intensity*, an important function of the observables? In figure 1, the x-axis for each plot shows the difference in means between treated and control on each of the covariates, as well as on *intensity*. All results are averaged over 500 simulations with the same data generating process and $N = 500$ on each simulation. The first plot (*matching*) shows results for simple Mahalanobis distance matching (with replacement). Imbalance remains somewhat large on *war duration*. More troubling, imbalance remains considerable on *intensity*, which was not directly included in the matching procedure. A careful researcher may realize the need to match on more functions of the covariates, and instead match on the original covariates, their squares, and their pairwise multiplicative interactions. While few researchers go this far in practice, the second plot in figure 1 (*matching+*) shows that even this approach would not provide the needed flexibility to produce balance on *intensity*. In fact, balance on both *war duration* and

intensity are worsened. In the third plot (*mean balance*), entropy balancing (Hainmueller 2012) is used to achieve equal means in the original covariates. As expected, this produces excellent balance on the original covariates, but only a modest improvement in balance on *intensity*. Finally, in the fourth plot (*kernel balance*), the kernel balancing approach introduced here is applied, again using the original covariate data alone. Because this method achieves balance on many smooth functions of the included covariates, it achieves vastly improved balance on *intensity*.

Figure 2: Biased ATT estimation due imbalanced function of the covariates



Boxplot illustrating distribution of average treatment effect on the treated (ATT) estimates in the same example as figure 1 above. The actual effect is zero *peace years*. *Matching*, *matching+*, and *mean balance* all show large biases because the control samples chosen by these procedures include higher *intensity* conflicts than the treated sample, even though *intensity* is entirely a function of observables. Since *intensity* influences the outcome, *peace years*, the treated and control samples thus differ regardless of any treatment effect. By contrast, *kernel balance* is approximately unbiased, as it achieves balance on a large space of smooth functions of the covariates.

These imbalances are worrying because they lead to biased ATT estimates: since *intensity* affects the outcome, mean differences in *intensity* between the treated and control group after adjustment lead to mean differences in the outcome between treated and control that are not due to the treatment. When the ATT is estimated by difference in means in the post weighting/matching sample, bias is thus found for all methods but kernel balancing (figure 2). While this example is artificial, it is reasonable in many cases that a function such as the ratio of two variables may impact the outcome variable in the absence of the treatment, yet investigators rarely ensure balance on such ratios. In general, it is unreasonable to expect investigators to correctly guess what functions of the observables may impact the outcome. Kernel balancing offers a solution to this problem.

3. Kernel Balancing

This section sets up the problem of ATT estimation, then describes the main ideas of the kernel balancing approach.

3.1. Setup

Using the Neyman-Rubin potential outcomes framework (see e.g. Rubin 1990; Sekhon 2008), let Y_{1i} and Y_{0i} be the treatment- and non-treatment potential outcomes respectively for units $i = 1, 2, \ldots, N$, and $D_i \in \{0, 1\}$ be the treatment assignment for unit i such that $D_i = 1$ for treated units and $D_i = 0$ for control units. The observed outcome for each unit is thus $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$. Suppose each unit has a vector of observed covariates, $X_i \in \mathbb{R}^P$, which cannot be affected by the treatment. Assume that Y_{1i}, Y_{0i}, X_i, and D_i are sampled independently from the joint density $p(X, Y_1, Y_0, D)$. We will be interested in the average treatment effect on the treated, $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$. However, when working with samples it will be more direct to consider the *sample average treatment effect on the treated* (SATT), $\hat{E}_N[Y_{1i} - Y_{0i}|D_i = 1]$, where $\hat{E}_N(\cdot)$ is the sample mean. This conditions on the sample in hand, though still includes the unobserved quantity $\hat{E}_N[Y_{0i}|D_i = 1]$. So long as the sample is drawn independently from $p(X, Y_1, Y_0, D)$, $\mathbb{E}[SATT] = ATT$, so estimates that are unbiased for the SATT will be unbiased for the ATT as well.

3.2. Bias of Difference in Means

Consider the (unweighted) difference in means estimand $DIM \equiv \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$, and its sample analog, $\widehat{DIM} \equiv \frac{1}{N_1}\sum_{i:D_i=1} Y_i - \frac{1}{N_0}\sum_{i:D_i=0} Y_i$ where N_0 is the number of control units and N_1 is the number of treated units.

The DIM is unbiased for the SATT (and the ATT) only when the treated and control groups would have the same expected outcome if neither had received the treatment. This allows the average outcome from the non-treated units to proxy for the average non-treatment potential outcome that the treated units would have had, had they not been treated. We can formalize this by decomposing the DIM estimator into the SATT and a bias:

$$\frac{1}{N_1}\sum_{i:D_i=1} Y_i - \frac{1}{N_0}\sum_{i:D_i=0} Y_i = \hat{\mathbb{E}}_N[Y_{1i}|D_i = 1] - \hat{\mathbb{E}}_N[Y_{0i}|D_i = 0] \tag{1}$$

$$= \hat{\mathbb{E}}_N[Y_{1i}|D_i = 1] - \hat{\mathbb{E}}_N[Y_{0i}|D_i = 1] + \hat{\mathbb{E}}_N[Y_{0i}|D_i = 1] - \hat{\mathbb{E}}_N[Y_{0i}|D_i = 0] \tag{2}$$

$$= \text{SATT} + \widehat{\text{Bias}} \tag{3}$$

It follows that the DIM is unbiased for the SATT simply when $\hat{\mathbb{E}}_N[Y_{0i}|D_i = 1] = \hat{\mathbb{E}}_N[Y_{1i}|D_i = 0]$, which I will refer to as "mean balance on Y_{0i}."

Lemma 1 (Mean Balance on Y_{0i} Implies Unbiasedness of DIM for SATT) *Provided the relevant moments exist, the difference in means (DIM) estimator is unbiased for the SATT if and only if mean balance on Y_{0i} holds, $\hat{\mathbb{E}}_N[Y_{0i}|D_i = 1] = \hat{\mathbb{E}}_N[Y_{0i}|D_i = 0]$.*

A final preliminary we need concerns identification assumptions. Just as matching, weighting, regresssion, and propensity score adjustment techniques, causal identification with kernel balancing will require that conditional ignorability holds. When identifying treament effects averaged only over treated units as in the ATT or SATT, this can be weakened slightly to conditional ignorability of the Y_{0i} alone,

Assumption 1 (Conditional Ignorability for the Non-treatment Outcome) *The non-treatment outcome is conditionally ignorable if*

$$Y_{0i} \perp\!\!\!\perp D_i | X_i$$

where Y_{0i} is the non-treatment potential outcome and is assumed to be bounded, D_i is treatment status, and X_i a vector of observed, pre-treatment covariates.

3.3. Obtaining Mean Balance on Y_{0i}

Kernel balancing differs from other methods in how it makes us of Assumption 1. Most matching and weighting methods can be understood as an effort to make the distribution of X_i the same for the treated and control (i.e. multivariate balance), after which a simple difference-in-means would be unbiaed for the ATT under Assumption 1. By contrast, kernel balancing targets the simpler goal of mean balance on Y_{0i}, allowing unbiased SATT estimation by Lemma 1. The *Discussion* section further explores the differences among methods.

Balance on Y_{0i} can be achieved without first ensuring multivariate balance on X, by imposing constraints on how $\mathbb{E}[Y_{0i}]$ relates to X_i. Specifically, for $X_i \in \mathbb{R}^N$, suppose that $\phi(\cdot)$ is some *feature expansion*, such that $\phi(X_i) : \mathbb{R}^P \mapsto \mathbb{R}^Q$ where Q may be (much) larger than N. For example, $\phi(X_i)$ could generate polynomial transformations of the original covariates. The specific nature of $\phi(X_i)$ used in kernel balancing relates to a Gaussian kernel, explained below. For the moment, the key feature of $\phi(X_i)$ needed here is that it is a sufficiently rich, non-linear expansion of X_i so that $\mathbb{E}[Y_{0i}|X_i]$ can be well fitted as a linear function of $\phi(X_i)$:

Assumption 2 (Linearity of Expected Non-treatment Outcome) *We assume that the conditional expectation of Y_{0i} is linear in the expanded features of X_i, $\phi(X_i)$, i.e.*

$$\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top \theta$$

for a feature expansion $\phi(\cdot) : \mathbb{R}^P \mapsto \mathbb{R}^Q$

We are interested in the choice of non-negative weights w_i on the control units that sum to 1, such that the weighted average vector $\phi(X_i)$ for the controls equals the unweighted average vector $\phi(X_i)$ for the treated,

Definition 1 (Mean balance on $\phi(X_i)$) *We say that w_i provides mean balance on $\phi(X_i)$ when:*

$$\frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) = \sum_{i:D_i=0} w_i \phi(X_i)$$

such that $\sum_i w_i = 1$, and $w_i \geq 0$ for all i.

Mean balance on $\phi(X_i)$ would be ideal to achieve because all linear functions of $\phi(X_i)$ would then have the same mean for the treated and control groups as well. To see this, note that assumption $\mathbb{E}[Y_{0i}|X_i]$ is linear in $\phi(X_i)$ (Assumption 2) is equivalent to assuming $Y_{0i} = \theta^\top \phi(X_i) + \epsilon_i$ where $\mathbb{E}[\epsilon_i|X_i] = 0$. We can then represent the sample mean of Y_{0i} for treated as

$$\frac{1}{N_1} \sum_{i:D_i=1} Y_{0i} = \frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i)^\top \theta + \epsilon_i \tag{4}$$

$$= \theta^\top \frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) + \frac{1}{N} \sum_{i:D_i=1} \epsilon_i \tag{5}$$

while the sample mean of Y_{0i} for the controls is

$$\sum_{i:D_i=0} w_i Y_{0i} = \sum_{i:D_i=0} w_i \left\{ \phi(X_i)^\top \theta + \epsilon_i \right\} \tag{6}$$

$$= \theta^\top \sum_{i:D_i=0} w_i \phi(X_i) + \sum_{i:D_i=0} w_i \epsilon_i \tag{7}$$

Recall that the bias of the SATT is the difference between the mean non-treatment potential outcomes for the treated and controls, $\theta^\top \frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) + \frac{1}{N} \sum_{i:D_i=1} \epsilon_i - \theta^\top \sum_{i:D_i=0} w_i \phi(X_i) + \sum_{i:D_i=0} w_i \epsilon_i$. Mean balance on $\phi(X_i)$ reduces this to $\frac{1}{N} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i$, which is zero in expectation. Mean balance on the linear bases $\phi(X_i)$ is thus sufficient for unbiased SATT estimation under Lemma 1. Note that the coefficients θ need not be determined.

If one had sufficient knowledge of $\mathbb{E}[Y_{0i}|X_i]$ to select a low-dimensional choice of $\phi(\cdot)$ while being confident that $\mathbb{E}[Y_{0i}|X_i]$ is linear in $\phi(X_i)$, then one could directly seek mean balance on each dimension of $\phi(X_i)$ and be confident that mean balance on Y_{0i} has been achieved. However, the general supposition of this paper is that investigators often have little knowledge of the functional form of $\mathbb{E}[Y_{0i}|X_i]$, except perhaps its continuity or expected smoothness. A very general choice of $\phi(X_i)$ is thus required, so that the functions $\phi(X_i)^\top \theta$ would include most reasonable functions. Yet, using a very high dimensional choice of $\phi(\cdot)$ (such as all squares, cubes, two-way, and three-way interactions, or indicators for deciles, etc.) to achieve this would make it difficult or impossible to achieve mean balance on $\phi(X_i)$.

3.4. Kernels

Fortunately, through a novel application of the "kernel trick", kernels allow us to side-step this problem, achieving mean balance on $\phi(X_i)$ even when it may be high- or infinite-dimensional.

For $X_i \in \mathbb{R}^P$, a kernel function, $k(\cdot, \cdot) : \mathbb{R}^P \times \mathbb{R}^P \mapsto \mathbb{R}$, takes in covariate vectors from any two observations and produces a single real-valued output interpretable as a measure of similarity between those two vectors. For reasons discussed below, we are interested principally in the Gaussian kernel:

$$k(X_j, X_i) = e^{-\frac{||X_j - X_i||^2}{s^2}} \tag{8}$$

Note that $k(X_i, X_j)$ produces values between 0 and 1 interpetable as a (symmetric) similarity measure, achieving a value close to 1 when X_i and X_j are most similar and approaching 0 as X_i and X_j become dissimilar. The choice parameter s might be called "scale", because it governs how close X_i and X_j must be in a Euclidean sense to be deemed similar. I discuss the choice of s further below. It is common to rescale each covariate prior to computing $k(X_i, X_j)$, dividing by the standard error. This ensures results will be invariant to unit-of-measure choices.

For a kernel that produces a positive semi-definite kernel matrix $\mathbf{K}$ (as a Gaussian kernel always does), there exists a choice of feature mapping $\phi(\cdot)$ such that $\langle \phi(X_i), \phi(X_j) \rangle = k(X_i, X_j)$. That is, for a given kernel, there exists a choice of expansion $\phi(\cdot)$ such that the inner-product of $\phi(X_i)$ and $\phi(X_j)$ can be computed by taking $k(X_i, X_j)$, even if $\phi(\cdot)$ cannot be explicitly formed.

The nature of $\phi(X)$ depends on the choice of kernel. For example, suppose $X_i = [X_i^{(1)}, X_i^{(2)}]$ and we choose the kernel $(1 + \langle X_i, X_j \rangle)^2$. This choice of kernel happens to corresponds to $\phi(X) = [1, \sqrt{2}X^{(1)}, \sqrt{2}X^{(2)}, X^{(1)}X^{(1)}, \sqrt{2}X^{(1)}X^{(2)}, X^{(2)}X^{(2)}]$, and one can confirm that $k(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$

for this choice of kernel and $\phi(\cdot)$. Using the Gaussian kernel, the corresponding $\phi(X)$ is infinite-dimensional. It suffices to note that this feature space has universal representation property: as $N \to \infty$, $\phi^\top\theta$ can fit any continuous function of X (Micchelli et al. 2006). Smoother functions can be fitted with fewer observations, making this an excellent choice to model $\mathbb{E}[Y_{0i}|X_i]$ when little is known about the nature of the relationship except that it is continuous and likely to be smooth. It is also a standard "workhorse" kernel for support vector machines, kernelized regressions, and other approaches where general functions need to be fitted. An intuition for the nature of this feature space (and the functions linear in it) is given in section 3.5.

Let $\mathbf{K}$ be the kernel matrix storing the results of each pairwise application of the kernel, i.e. $\mathbf{K}_{\{i,j\}} = k(X_i, X_j) = \langle\phi(X_i)\phi(X_j)\rangle$. To reduce notation it is useful to order the observations so that the N_1 treated units come first, followed by the N_0 control units. Then $\mathbf{K}$ can be partitioned into two rectangular matrices,

$$\mathbf{K} = \begin{bmatrix} \mathbf{K_t} \\ \mathbf{K_c} \end{bmatrix}$$

where $\mathbf{K_t}$ is $N_1 \times N$ and $\mathbf{K_c}$ is $N_0 \times N$. The average row of $\mathbf{K}$ for the treated can then be written $\frac{1}{N_t}\mathbf{K_t}\mathbf{1}_{N_t}$, while the weighted average row of $\mathbf{K}$ is $\mathbf{K_c}w$ for the $N_0 \times 1$ vector of weights w, with weights summing to 1.

3.5. Mean balance on $\mathbf{K}$

Working with kernels and constructing the kernel matrix $\mathbf{K}$ pays off because mean balance on $\phi(X)$ is achieved by getting mean balance on the columns of $\mathbf{K}$. Consider a single row of $\mathbf{K}$:

$$k_i = [k(X_i, X_1), k(X_i, X_2), \ldots, k(X_i, X_N)]$$

which describes each observation not in terms of it's original X coordinates but as a vector of N similarities to each of the observations. Similar to mean balancing on X_i, kernel balancing then seeks weights that ensure the average k_i of the treated is equal to the weighted mean vector k_i of the controls:

Definition 2 (Mean balance on $\mathbf{K}$) *The weights w_i achieve mean balance on $\mathbf{K}$ when*

$$\overline{k_t} = \sum_{i:D=0} w_i k_i$$

such that $\sum_i w_i = 1$, and $w_i \geq 0$ for all i, where $\overline{k_t}$ is the average row of $\mathbf{K}$.

This achieves mean balance on the corresponding $\phi(X_i)$ even if $\phi(X_i)$ has dimensionality (much) greater than N.

Proposition 1 (Balance in $\mathbf{K}$ implies balance in $\phi(X)$) *Let the mean row of $\mathbf{K}$ among the treated units be given by $\overline{k_t} = \frac{1}{N_t}\mathbf{K_t}\mathbf{1}_{N_t}$ and the weighted mean row of $\mathbf{K}$ among the controls given by $\mathbf{K_c}w$. If $\overline{k_t} = \mathbf{K_c}w$, then $\overline{\phi_t} = \overline{\phi_c}$ where $\overline{\phi_t} = \frac{1}{N_t}\sum_{D_i=1}\phi(X_i)$ and $\overline{\phi_c} = \sum_{D_i=0}\phi(X_i)w_i$.*

Proposition 1 is a novel application of the "kernel trick" and implies that the treated and control groups have the same mean on each dimension of $\phi(X)$ when the rows of $\mathbf{K}$ for the treated and control have the same means, regardless of the dimensionality of $\phi(\cdot)$. Proof is given in the appendix.

Finally, the weights w_i that produce mean balance on $\mathbf{K}$ in a finite sample can be used in a difference in means estimation. The main result can now be stated:

THEOREM 1 (UNBIASEDNESS OF WEIGHTED DIFFERENCE IN MEANS FOR THE SATT) *Consider the weighted difference in means estimator,*

$$\widehat{DIM}_w = \frac{1}{N}\sum_{i:D_i=1} Y_i - \sum_{i:D_i=0} w_i Y_i$$

$$\textit{such that } \overline{k_t} = \sum_{i:D=0} w_i k_i,\ \sum_i w_i = 1 \textit{ and } w_i > 0$$

Under assumptions of conditional ignorability for the non-treatment outcome (Assumption 1) and linearity of $\mathbb{E}[Y_{0i}|X_i]$ *in* $\phi(X_i)$ *(Assumption 2),* $\widehat{DIM}_w$ *is unbiased for the sample average treatment effect on the treated (SATT) and the (population) ATT.*

Proof is given in the appendix (7.1), though the intuition is simple and helps to summarize the approach: mean balance in $\mathbf{K}$ gives mean balance in $\phi(X)$, which produces mean balance for functions linear in $\phi(X)$, including the conditional expectation of Y_{i0}. The appendix also describes the bias under conditions in which $\mathbb{E}[Y_{0i}|X_i]$ lies outside the span of $\phi(X_i)$.

3.6. Implementation

What remains is to choose the weights, w_i, to obtain mean balance on $\mathbf{K}$. This section describes one possible estimation procedure, as implemented in the `R` package, `kbal`.

A method is needed to find the weight vector w such that $\frac{1}{N_1}\mathbf{K}_t\mathbf{1}_{N_1} = \mathbf{K}_c w$, while constraining the weights to be non-negative and sum to one. It is also desirable to do this with minimal variation in the weights, as measured by some metric. Two natural candidates for this are empirical likelihood (Owen 1988), and entropy balancing (Hainmueller 2012), though other approaches such as those that explicitly minimize the variation in weights for a given degree of imbalance (e.g. Zubizarreta 2015) may be valuable as well. Here I employ entropy balancing, which seeks to satisfy these conditions while maximizing the entropy implied by the weights.

Establishing balance on all columns of $\mathbf{K}$ is typically infeasible, owing to the near co-linearity of many columns of $\mathbf{K}$, and this co-linearity is perfect in cases where a single observation is repeated exactly. Thus, a dimension reduction step is needed. One option would begin with a supervised learning approach: we are only interested in obtaining mean balance on Y_{0i}, and so truly only need mean balance on columns of $\mathbf{K}$ found to predict of Y_{0i}. For example, a lasso (Tibshirani 1996) regression of Y_{0i} on X_i (using only control units) could be used to choose which columns of $\mathbf{K}$ should be balanced. This approach has potential virtues in terms of efficiency. However, (a) this makes results dependent upon an additional estimation, and (b) some investigators prefer using balancing procedures that do not utilize the outcome data (Rubin 2007) to maximize transparency. While potentially promising, I leave this approach for future work, and focus here on choosing dimensions of $\mathbf{K}$ using unsupervised dimension reduction strategies.

Here, I take an unsupervised approach to dimension reduction, projecting $\mathbf{K}$ onto its major principal components. Projections of $\mathbf{K}$ corresponding to the largest eigenvalues are selected for balancing first. The optimization procedure is then over the parameter, *numdims*, which controls how many of the "most important" projections of $\mathbf{K}$ are balanced. Since the goal is to achieve $\overline{k_t} = \sum_{i:D=0} w_i k_i$, a natural loss function to judge the success of a set of weights would be $a||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ for some norm $||\cdot||$ and constant a. One reasonable choice is the L_1 norm, using

$$L_1 = \frac{1}{2}\sum_{i=1}^{N} |\overline{k_t} - \sum_{i:D=0} w_i k_i|^1 \tag{9}$$

While motivated as a measure of imbalance on $\mathbf{K}$, we will see that this quantity is also a measure of the difference between the multivariate distribution of the covariates for the treated and control, when those distributions are estimated using a certain smoother (see *Discussion*). For the optimization, the first *numdim* projections of $\mathbf{K}$ are kept and weights are chosen to ensure mean balance on them. The parameter *numdims* is then varied, until the L_1 imbalance measure is minimized. Typically, imbalance measured by L_1 improves as *numdims* initially rises, and then deteriorates once *numdims* is too high and numerical instability begins to creep in. An illustration of the relationship *numdims*, L_1 and the balance achieved on unknown functions of X is given in the appendix (figure 6)

4. Discussion

Having described the basic logic and procedure for kernel balancing, I now remark on its relationship to existing procedures, some additional properties and implications of this approach, and further implementation details.

4.1. Relation to Existing Balance Approaches

Here, I compare kernel balancing to matching, covariate balancing weights, and propensity score methods. Like kernel balancing, each of these begins with a conditional ignorability assumption (Assumption 1). However they exploit this assumption to make causal inferences through a more difficult estimation route of seeking multivariate balance.

Matching

Under conditional ignorability as defined in Assumption 1, treatment assignment is independent of potential outcomes within each stratum of X. The most natural way to exploit this for estimating the SATT is to perform this conditioning on X very literally: take difference-in-means estimates of the treatment effect within each stratum of X, then average these together over the empirical distribution of X for the treated. Subclassification and exact matching estimators do this. However, conditioning on X in this way is impractical or impossible when X is continuous or contains indicators for many categories, since we cannot literally compute differ tences for each stratum of X.

Matching approaches (e.g. Rubin 1973) mimic this conditioning, taking each treated unit in turn, finding the nearest one or several control units, and retaining only these control units in the sample (typically with replacement). A difference-in-means on the outcomes in the resulting matched data is the same as an average over the differences within each pairing. The method works

when multivariate balance is achieved through the matching procedure, i.e. the distribution of X for the control units becomes the same as the distribution for the treated units. The non-parametric nature of matching is appealing as a multivariate balancing technique, but its accuracy is limited by the problem of matching discrepancies. Specifically, in a given pairing, the treated unit may be systematically different on X than the control unit(s) it is paired with when exact matches cannot be found. Thus the conditioning on X is incomplete, and the distribution of X for the treated and controls are not identical. The bias in (S)ATT estimates this causes dissipates only very slowly as we average over more matched groups, and in general the resulting estimates are not $\sqrt{N}$-consistent (Abadie and Imbens 2006). Investigators are instructed to try different matching approaches until they achieve satisfactory multivariate balance (see e.g. Stuart 2010). However in practice, multivariate densities are difficult to measure or compare, so tests for this balance are usually limited to univariate tests comparing the marginal distribution of each covariate under treatment and control. In short, the goal of matching is to align the multivariate distribution of covariates for the control units with that of the treated, but matching discrepancies can prevent this from occurring, and the tools used to test for this multivariate balance are incomplete. As the motivating example in section 2 illustrates, matching can thus fail even when investigators attempt to match on higher-order terms.

Covariate Balancing Weights

Another category of methods for multivariate balancing are covariate balancing weighting techniques that use probability-like weights on the control units to achieve a set of prescribed moment conditions on the distribution of the covariates among the controls (e.g. univariate means and variances). Examples from the causal inference literature include entropy balancing (Hainmueller 2012) and the covariate balancing propensity score (Imai and Ratkovic 2014), with a number of similar procedures emerging from the survey sampling literature, such as raking (Kalton 1983). Once these moment conditions are satisfied, it is assumed that the multivariate densities for the treated and control are alike in all important respects. These weights can be used in a difference in means estimation or other procedure. The upside of this procedure over matching is that the prescribed moments of the control distribution can be made exactly equal to those of the treated, avoiding the matching discrepancy problem. The downside is that it loses the non-parametric quality of matching, providing balance only on enumerated moments. It is generally not possible to know what moments of the distribution must be balanced to ensure unbiasedness, because we do not know which functions of the covariates might influence the (non-treatment) outcome. Kernel balancing can be understood as an extension to these covariate balancing weighting methods that solves this problem by ensuring balance on a large class of functions of the covariates automatically.

Propensity Score Weighting

Finally, propensity score methods such as inverse propensity score weighting are similarly an attempt to find the weights that make the distribution of the covariates for the controls and treated similar, through adjusting for treatment probabilities. It is useful to show more explicitly the role played by inverse propensity score weights in estimating the ATT as this will be used again in describing further properties of kernel balancing below.

Under Assumption 1, the ATT can be re-written:

$$ATT = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \tag{10}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 1)dx \tag{11}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 0, x]p(x|D_i = 1)dx \tag{12}$$

Expression 12 is identifiable in the sense that we only require treatment potential outcomes from the treated units, and non-treatment potential outcomes from the non-treated units. However, it remains problematic because it requires averaging outcomes from control units over the distribution of X for the treated, $p(x|D_i = 1)$, which is not the distribution of the control units in the sample. Specifically, the difference in means estimand,

$$DIM = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \tag{13}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 0)dx \tag{14}$$

differs from the ATT in its second term, because it averages over the outcomes of non-treated units at their natural density in X, $p(x|D_i = 0)$. To address this, consider a weighted difference in means estimand,

$$DIM_w = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}_w[Y_{0i}|D_i = 0] \tag{15}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int w_i\mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 0)dx \tag{16}$$

where w_i is a function of X that allows us to upweight or downweight control units. The difference between expression 12 and 14 can be resolved by choosing weights

$$w_i = \frac{p(x|D_i = 1)}{p(x|D_i = 0)} \tag{17}$$

Through Bayes theorem, we can replace the class densities in this expression with more familiar propensity scores to obtain $w_i = \frac{p(D_i=1|x)p(D_i=0)}{p(D_i=0|x)p(D_i=1)}$. For the control units ($D_i = 0$), this is $w_i = \frac{p(D_i)}{p(D_i|X_i)}\frac{1-p(D_i|X_i)}{1-p(D_i)}$. These are the stabilized inverse propensity scores one would apply to the control units to estimate the ATT. These weights, if properly estimated, ensure that the whole distribution of X for the control units is adjusted to equal the distribution among the treated.

Comparing these methods, note that achieving multivariate balance through any of these tools asks a great deal in terms of the number of effective quantities that must be made equal for the treated and control groups (the density at each location in the covariate space) and the high variance and instability of resulting solutions. A key feature of kernel balancing, by contrast, is that it targets the "mean balance in Y_{0i}" goal instead, which is all that is required for unbiased estimation of the SATT (Lemma 1). It does so at the cost of an assumption that $\mathbb{E}[Y_{0i}|X_i]$ is linear in some basis functions $\phi(X_i)$. Through the use of kernels, however, this assumption can be made very mild as $\mathbb{E}[Y_{i0}|X_i]$ need only be smooth (or for large enough N, at least continuous) in X_i. That said, this procedure also has useful implications for achieving multivariate balance, discussed below (section 4.3).

4.2. Functions of X that can be balanced

In kernel balancing, the only function of the covariates that must have the same mean for treated and controls is $\mathbb{E}[Y_{0i}|X_i]$. An equivalent view that is useful in thinking about the need for balance is that any function of the covariates that has a different mean for the treated and control group can cause bias in the SATT estimate if it also influences the non-treatment potential outcome. Either way, kernel balancing attempts to obtain mean balance on these functions, and does so if they are well fitted by $\phi(X_i)^\top\theta$ for some θ, where $\phi(X_i)$ is induced by the choice of Gaussian kernel.

Understanding what this function space looks like can be difficult, since the choice of $\phi(X_i)$ implied by the Gaussian kernel is infinite-dimensional. Nevertheless, in a given finite sample, an easy and accurate intuition is available. The functions linear in $\phi(X_i)$ are those that can be built from the superposition of Gaussians placed over each observation and arbitrarily rescaled. That is, suppose we place a Gaussian kernel over each observation in the dataset, rescale each of these Gaussians by a value of c_i for that observation, then sum the resulting rescaled Gaussians to form a single surface. By varying the scaling factors in c, an enormous variety of smooth functions can be formed in this way, approximating a wide variety of non-linear functions of the covariates. This view is described and illustrated at length in Hainmueller and Hazlett (2014), where this function space is used to model highly non-linear but smooth functions even in high-dimensional problems and with relatively few observations. This space of functions is appealing because while making no assumptions of linearity or additivity in X, it is generally reasonable to assume that the conditional expectation of Y_{0i} is continuous and relatively smooth in X.

Kernel balancing thus provides balance on many functions of the covariates, with smoother functions being the most easily balanced in small samples. It does this without having to specify exactly what functions of the covariates we wish to ensure balance upon. As shown in the motivating example (section 2), this ensures that various reasonable functions of the covariates that the user does not know to check balance on will have equal means for the treated and controls.

4.3. Smoothed multivariate balance

The principle goal of kernel balancing is obtaining mean balance on the non-treatment potential outcomes, however this procedure does imply that a particular *estimate* of the multivariate density of the covariates is equal for the treated and control groups at all locations in the dataset. It thus helps to achieve the goals normally targeted by matching and weighting procedures.

Proposition 2 (Balance in **K** implies equality of smoothed multivariate densities) *Consider a density estimator for the treated, $\hat{f}_{X|D=1}$ and for the (weighted) controls, $\hat{f}_{X|D=0,w}$, each constructed with kernel $k(\cdot,\cdot)$ of bandwidth s^2 as described below. The choice of weights that ensures mean balance in the kernel matrix* **K** *ensures that $\hat{f}_{X|D=1} = \hat{f}_{X|D=0,w}$ at every position at which an observation is located.*

Proof of proposition 2 is given in the appendix. Here I briefly build an intuition for this result, as it leads to further considerations. First, the typical Parzen-Rosenblatt window approach estimates a density function according to:

$$\hat{f}(x) = \frac{1}{N\sqrt{\pi s^2}} \sum_{i=1}^{N} k(x, X_i) \tag{18}$$

for kernel function $k(\cdot, \cdot)$ with bandwidth s^2, where the normalizing constants are required since they are not included in the definition of the Gaussian kernel used throughout this paper.

The Gaussian kernel is among the most commonly used for this task. While typically considered in a univariate context, expression 18 utilizing a Gaussian kernel generalizes to a multivariate density estimator based on Euclidean distances. Such density estimators are intuitively understandable as a process of placing a multivariate Gaussian kernel over each observation's location in $\mathbb{R}^P$, then summing them into a single surface and rescaling, providing a density estimate at each location.

The link between obtaining mean balance on Y_{0i} and obtaining multivariate density balancing emerges from the fact that both are manipulations of the superpositions of kernels placed over each observation. For a sample consisting of $X_1, \ldots, X_N$, construction of the kernel matrix $\mathbf{K}$ using the Gaussian kernel and right-multiplying it by a column vector, $\frac{1}{N\sqrt{\pi s^2}}$, produces values numerically equal to first constructing such an estimator based on all the observations represented in the columns of $\mathbf{K}$, then evaluating the resulting density estimates *at all the positions represented by the rows of* $\mathbf{K}$. To see this, consider that the value of $\mathbf{K}a$ at a given point X_j is $\sum_i a_i k(X_i, X_j)$. Note that $k(X_i, X_j)$ is the value that would be obtained by placing a Gaussian over X_i and evaluating its height at X_j. Thus $\sum_i a_i k(X_i, X_j)$ is the value that would be obtained by placing a Gaussian kernel over each observation, X_i, and evaluating the height of the resulting summated surface at X_j. Similarly, the expression $\frac{1}{N_1\sqrt{\pi s^2}}\mathbf{K_t}^\top \mathbf{1}_{N_t}$ where $\mathbf{1}_{N_t}$ is a N_t-vector of ones thus returns a vector of estimates for the density of the treated, as measured at all observations. Likewise, $\frac{1}{N_0\sqrt{\pi s^2}}\mathbf{K_c}^\top \mathbf{1}_{N_0}$ returns estimates for the density of the control units at every datapoint in the sample, and $\frac{1}{\sqrt{\pi s^2}}\mathbf{K_c}^\top w$ gives the w-weighted density of the controls, again as measured at every observation.

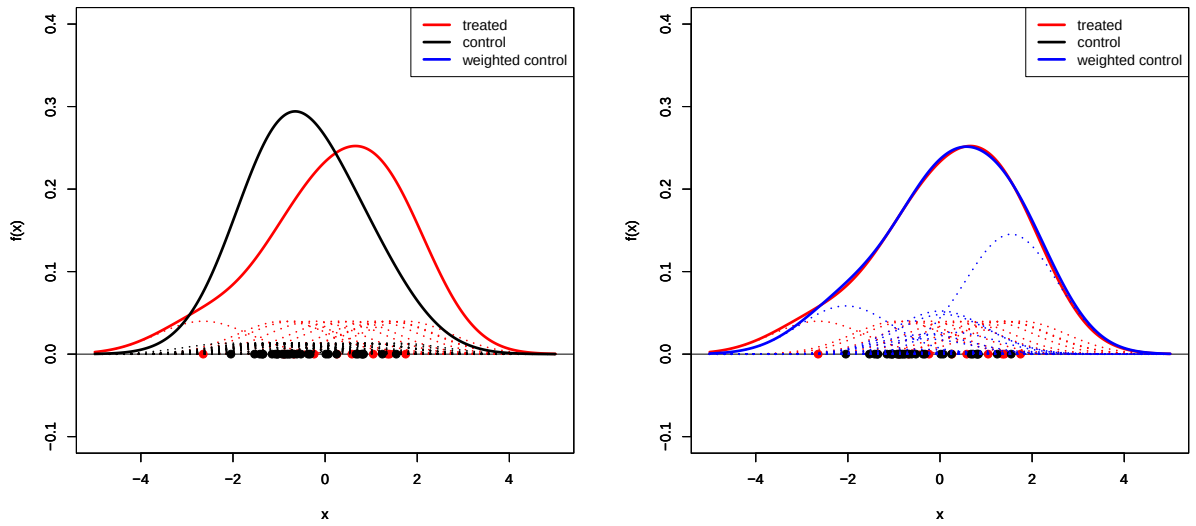
If we take these estimates as reasonable measures of density, we would like to choose the weights such that the weighted density of the controls equals that of the treated, at every observation. Proposition 2 states that the choice of w found by kernel balancing to achieve $\mathbf{K}_c^\top w = \frac{1}{N_1}\mathbf{K}_T^\top \mathbf{1}_{N_1}$ is exactly the choice that equalizes these smoothed density estimates for the treated and weighted controls at every point in the dataset. Proof is given in the appendix.

Figure 3 provides a graphical illustration of the density-equalizing property of the kernel balancing weights for a one-dimensional problem. This density equalizing view connects kernel balancing more directly to other approaches such as matching, but it is important to remember that it is mean balance in Y_{0i} that is essential for unbiasedness, and which kernel balancing targets. Kernel balancing only equalizes the densities *as they are estimated* by the smoothing action of the selected kernel. In some cases a density estimate constructed in this way would not be a natural one, for example when X is a categorical variable or has sharp bounds. Nevertheless, this approach will apply the same smoothing estimator to the treated and to the control.

4.4. Further Implications of the Multivariate Balance Property

If the measure of multivariate density given by the kernel approximation is satisfactory, a number of further interpretations and links to other methods are made apparent. One implication is that the kernel balancing weights are marginalizing weights, i.e. those that make $p(D|X) = p(D)$. This makes it possible to run any number of analyses on the weighted data ignoring X, instead using the weights. A second implication is that this approach automatically achieves what any propensity score method would hope to. By making $p(D_i = 1|X_i) = p(D_i)$ and $p(D_i = 0|X_i) = p(D_i = 0)$, the weights chosen here are equivalent to stabilized inverse propensity score weights as in 17, since these weights would simply equal 1 after weighting by w_i. Critically, each of these statements can

Figure 3: Density Equalizing Property of the *kbal* Weights



Left: Density estimates for treated and (unweighted) controls. Red dots show the location of 10 treated units. Dashed lines show the appropriately scaled Gaussian over each observation, which sum to form the density estimator for the treated (red line) and control (black line). The L_1 imbalance (see below) is measured to be 0.32. *Right:* Weights chosen by kernel balancing effectively rescale the height of the Gaussian over each control observation (dashed blue lines). The new density estimate for the weighted controls (solid blue line) now closely matches the density of the treated at each point. The L_1 imbalance is now measured to be 0.002

only be made if we take the particular density estimate implied by the choice of kernel as correct.

4.5. The L_1 measure of both K-imbalance and multivariate density imbalance

As discussed above, the optimization procedure chooses the number of projections of $\mathbf{K}$ that must be balanced while seeking to minimize overall imbalance on $\mathbf{K}$. Minimizing an imbalance measure of the form $a||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ for some norm $||\cdot||$ is natural given the goal of mean balance on $\mathbf{K}$.

Such a norm also provides a measure of continuous multivariate imbalance. Setting a to $\frac{1}{\sqrt{\pi s^2}}$ to obtain $||\frac{1}{N_1\sqrt{\pi s^2}}\mathbf{K_t}^\top \mathbf{1}_{N_1} - \frac{1}{\sqrt{\pi s^2}}\mathbf{K_c}^\top w||$ we see this equals $||\hat{f}_{D=1}(\mathbf{X}) - \hat{f}_{w,D=0}(\mathbf{X})||$, a norm on the difference between the smoothed density estimators for the treated and (weighted) controls, evaluated at each observation in the dataset (see figure 3). Hence the proposed quantity is a natural one not only from the perspective of a loss function on our estimation target, but also for minimizing the difference between the smoothed density estimates for the treated and weighted controls. When interpreted as a difference between estimated densities, this metric is similar to the L_1 metric used in Coarsened Exact Matching (Iacus et al. 2011), but without requiring coarsening in order to construct discrete bins.

4.6. Choice of s^2

Since mean balance on Y_{0i} is the primary goal, not density estimation or equalization, the choice of the kernel and s should be made accordingly. While it is tempting to think of s^2 as the usual bandwidth that must be carefully selected in density estimation procedures, here it is much more

important to choose s according to how it effects mean balance in Y_{0i}. To this end, the choice of parameter s^2 is a feature-extraction decision that determines the construction of $\phi(X_i)$ and thus $\mathbf{K}$. It determines how close two points X_i and X_j need to be in order to have highly similar rows k_i and k_j. This implies a bias-variance tradeoff. If s^2 is too large, mean balance is easier to achieve and the weights will have low variance, the resulting balance is less precise (and the correspondding smoothed densities more "blurred"). If s^2 is too small, $\mathbf{K}$ will approximate the identity matrix, and each row k_i will be nearly linearly independent. In this case, the algorithm will not converge as balance cannot be attained. (The possibility of trimming away treated units that are difficult to match under small s^2 is discussed in Appendix 7.5).

Fortunately, in many cases balance is achievable across a wide range of s^2 values, and estimated SATTs are stable across a wide range. While lower values of s^2 are generally preferable, they risk higher variance, potentially placing large weights on a small proportion of the controls. For an easily interpretable metric, I propose the quantity *min90*, which is the minimum number of control units that are required to account for 90% of the total weight among the controls. For example, if *min90*=20, 90% of the total weight of the controls comes from just the 20 most heavily-weighted observations. This gives the user a sense of how many control units are effectively being used. The empirical example below shows how this can be used.

I propose choosing $s^2 = 2dim(X)$ as a reporting standard, while showing results at other choices for robustness. The square of the average Euclidean distance $\mathbb{E}[||X_i - X_j||]$ in the kernel calculation scales with $dim(X)$. Choosing s^2 proportional to $dim(X)$ thus ensures a relatively sound scaling of the data, such that some observations appear to be closer together, some further apart, and some in-between, regardless of $dim(X)$. A similar logic has been proposed for regression technique using a Gaussian kernel (see e.g. Hainmueller and Hazlett 2014; Schölkopf and Smola 2002). The constant of proportionality, however, remains open to debate. Empirically, the choice of $s^2 = 2dim(X)$ has offered very good performance, and so this is the default value of s^2, though clearly further work is needed to justify this choice. Though results are often not sensitive to the choice of s^2, investigators may wish to present their results across a range of s^2 values to ensure this choice is not consequentially in a given application. Where results do vary across s^2 values, inspecting L_1 and $min90$ can be helpful for determining an appropriate value.

5. Empirical Examples

It is useful to know whether kernel balancing accurately recovers average treatment effects in observational data under conditions in which an approximately "true" answer is known. This can be approximated using a method and dataset first used by LaLonde (1986) and Dehejia and Wahba (1999), and which has become a routine benchmark for new matching and weighting approaches (e.g. Diamond and Sekhon 2005; Iacus et al. 2011; Hainmueller 2012).

The aim of these studies is to recover an experimental estimate of the effect of a job training program, the National Supported Work (NSW) program. Following LaLonde (1986), the treated sample from the experimental study is compared to a control sample drawn from a separate, observational sample. Methods of adjustment are tested to see if they accurately recover the treatment effect despite large observable differences between the control sample and the treated sample.See Diamond and Sekhon 2005 for an extensive description of this dataset and the various subsets that have been drawn from it. Here I use 185 treated units from NSW, originally selected by Dehejia and Wahba (1999) for the treated sample. The experimental benchmark for this group of treated

units is \$1794, which is computed by difference-in-means in the original experimental data with these 185 treated units. The control sample is drawn from the Panel Study of Income Dynamics (PSID-1), containing 2490 controls.

The pre-treatment covariates available for matching are age, years of education, real earnings in 1974, real earnings in 1975 and a series of indicator variables: Black, Hispanic, and married. Three further variables that are actually transforms of these are commonly used as well: indicators for being unemployed (having income of \$0) in 1974 and 1975, and an indicator for having no highschool degree (fewer than 12 years of education).

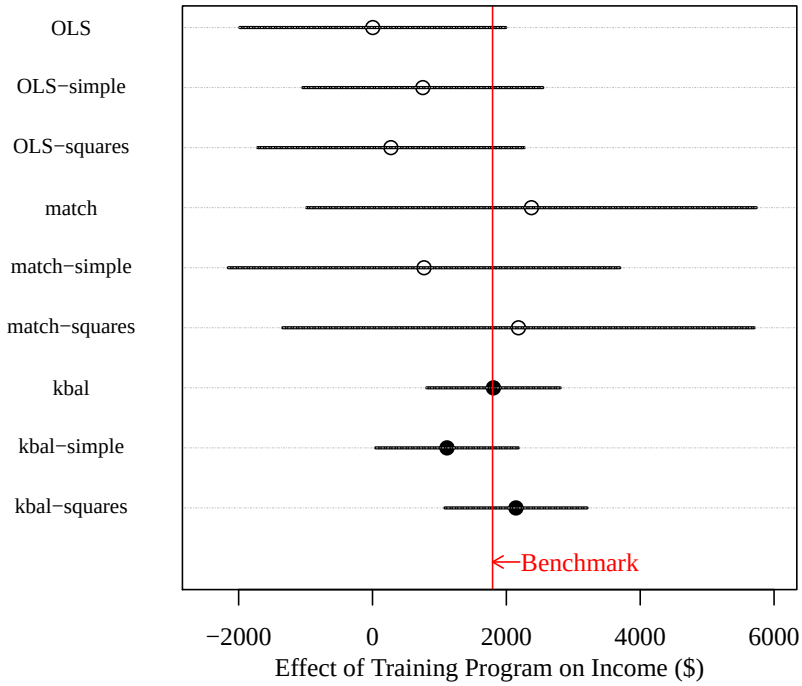
As found by Dehejia and Wahba (1999), propensity score matching can be effective in recovering reasonable estimates of the ATT, but these results are highly sensitive to specification choices in constructing the propensity score model (Smith and Todd 2001). Diamond and Sekhon (2005) use genetic matching to estimate treatment effects with the same treated sample. While matching solutions with the highest degree of balance produced estimates very close to the experimental benchmark, these models included the addition of squared terms and two-way interactions. Similarly, entropy balancing Hainmueller (2012) has also been shown to recover good estimates using a similar setup, using a control dataset based on the Current Population Survey (CPS-1), employing all pairwise interactions and squared terms for continuous variables, amounting to 52 covariates.

Figure 4 reports results from a variety of estimation procedures and specifications. Three procedures are used: linear regression (*OLS*), Mahalanobis distance matching (*match*), and kernel balancing (*kbal*). For *match* and *kbal*, estimate are produced by simple difference in means on the matched/reweighted sample. Standard errors from matching are the Abadie-Imbens standard errors. Standard errors for kernel balancing are from weighted least squares with fixed weights. Bootstrapped standard errors may be preferable for kernel balancing to account for uncertainty in the choice of weights.

For each method, three sets of covariates are attempted: the standard set of 10 covariates described above, a reduced set (*simple*) including only the seven of these that are not transforms of other variables, and an expanded set (*squares*) including the 10 standard covariates plus squares of the three continuous variables. Figure 4 shows that the OLS estimates vary widely by specification, and even the estimate closest to the benchmark (\$1794) is incorrect by \$1042. Mahalahobis distance matching performs better, though remains somewhat specification dependent, with its best estimate (*match-squares*) falling within \$387 of the benchmark. Finally, kernel balancing performs well over the three specification. While there is some variation by specification, no estimate is more than \$681 from the benchmark, and the standard specification, *kbal*, produces an estimate of \$1807, within \$13 (0.7%) of the benchmark.

From the kernel balancing solution, we can also see that balance is difficult to achieve in this example, in the sense that it requires focusing on a relatively small portion of the original control sample. Specifically, at the solution achieved by kernel balancing, $min90 = 193$, meaning that 90% of the total weight of the control comes from 193 observations. While this is still a reasonable number, and similar to the size of the treatment group, it implies that approximately 90% of the control sample was not useful for comparison to the treated. This is appropriate, however, given the large differences between the treated and control samples. For example, while 72% of the treated are unemployed in either 1974 or 1975, only 12% of controls are unemployed in either year.

Figure 4: Estimating the Effect of a Job Training Program from Partially Observational Data



Reanalysis of Dehejia and Wahba (1999), estimating the effect of a job training program on income. Three procedures are used: linear regression (*OLS*), Mahalanobis distance matching (*Match*), and kernel balancing (*kbal*). For each, three sets of covariates are attempted: the standard set of 10 covariates described in the text, a reduced set (*simple*) including only the seven of these that are not transforms of other variables, and an expanded set (*squares*) including the 10 standard covariates plus squares of the three continuous variables. While *OLS* and *match* perform reasonably well, both are sensitive to specification. The best OLS estimate (*OLS-simple*) still under-estimates the $1794 benchmark by $1042, while the best matching estimate (*match-squares*) is off by $387. Kernel balancing performs reasonably well on all three specification, and the standard specification, *kbal*, produces an estimate of $1807, within $13 of the benchmark.

5.1. Are Democracies Inferior Counterinsurgents?

The second example, found in the Appendix (7.6), applies kernel balancing to a re-examine whether democracies are less successful in fighting counterinsurgencies (Lyall 2010). While theory and prior research has argued found that democracies are inferior counterinsurgents, Lyall (2010) finds no such relationship, using a novel dataset and matching to ensure comparability of the counterinsurgencies fought by democracies and non-democracies. Reexamining the post-1945 period and using the same covariates, kernel balancing proves far more effective in obtaining balance than the original matching procedure, both on the covariates directly included in the balancing procedures, and on functions of these variables. Using five different models to estimate the effect of democracy on the adjusted datasets, estimates from the kernel balanced data all indicate that democracies were 26 to 27 percentage points less likely to win counterinsurgencies over this period than non-democracies, in contrast to the null finding of (Lyall 2010). These effects are statistically significant, but also substantively large, especially given the overall success rate of just 33%.

6. Conclusions

In the ongoing quest to reliably infer causal quantities from observational data, the first-order challenge often remains ensuring that there are no unobserved confounders in a given identification scenario, so that assumptions such as Assumption 1 are plausible. However even under these assumptions, the actual problem of conditioning on observables to estimate causal effects remains non-trivial. Matching, covariate balancing weights, and propensity score weighting each seek to make the multivariate distribution of covariates for the untreated equal to that of the treated. If any function of the observables that influences the non-treatment outcome persists in having a different mean for the treated and controls, the resulting estimates will be biased. Unfortunately, the investigator is not generally aware of all the functions of the covariates that may influence the outcome, making it difficult to guard against this possibility.

However, unbiasedly estimating the SATT requires only that $\hat{\mathbb{E}}_N[Y_{0i}|D_i = 1] = \hat{\mathbb{E}}_N[Y_{0i}|D_i = 0]$, or "mean balance on Y_{0i}". Kernel balancing achieves this goal by working with the kernel matrix, $\mathbf{K}$, rather than the original covariates, X. It finds weights on the controls to make the weighted average row of $\mathbf{K}$ for the controls equal to the average row of $\mathbf{K}$ for the treated. Mean balance on these features implies mean balance on all functions that can be formed by the superposition of Gaussians placed over each observation in the covariate space. The assumption that $\mathbb{E}[Y_{0i}|X_i]$ is among these functions is far more plausible than the assumption that it is linear in the original X, even if the investigator is careful enough to include higher-order terms among these X's. Moreover as N grows large, $\mathbb{E}[Y_{0i}|X_i]$ is increasingly well modeled within this space.

While mean balance on Y_{0i} is the principle goal, kernel balancing also implies that a particular kernel-based smoother for the multivariate densities is equal for the treated and control, at every observations. Insofar as this is a reasonable density estimate, kernel balancing thus achieves what matching and covariate balancing estimators seek to achieve. These weights are also equivalent to a stabilized inverse propensity score weight that does not require an explicit model for the propensity score, and weights that achieves multivariate balance as measured by a kernel smoother. This smoothed multivariate balance is achieved in a given sample, not just in expectation as is the case with traditional propensity score estimation. Thus, while focusing first on the minimum requirement for unbiased SATT estimation, the method also achieves the goals for which matching, weighting, and propensity score have traditionally been employed.

Kernel balancing performs well in a reanalysis of Dehejia and Wahba (1999), a widely used benchmark for covariate adjustment in causal inference. At it's default values, with the covariates commonly used for this problem and no further specification choices, kernel balancing estimated an effect of \$1807 using the non-experimental control group, extremely close to the experimental benchmark of \$1794.

Numerous questions and challenges remain for future work. First, it will be useful to better understand the asymptotic properties of this procedure and its comparative performance relative to other approaches when the expectation of Y_{0i} is non-linear in X_i. Second, $\mathbf{K}$ has dimensionality $N \times N$, which becomes unwieldy as N grows large, posing a practical limit of tens of thousands of observations. Third, obtaining correct confidence intervals for estimates based on weighted samples – either through resampling or a closed-form solution – will be an important and useful advance, particularly since standard errors remain poorly understand for matching techniques. Fourth, better results may be obtained by a procedure that selects what features of $\phi(X)$ or what columns of $\mathbf{K}$ influence Y_{0i} and seek balance preferentially on these rather. Finally, improvements may be

possible on a number of further implementation details, such as the choice of s^2, the optimization procedure for choosing the number of dimensions, alternate methods for dimension reduction on $\mathbf{K}$, and alternative methods for choosing the balancing weights that achieve mean balance on $\mathbf{K}$ while minimizing volatility.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica* **74**(1): 235–267.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American statistical Association* **94**(448): 1053–1062.

Diamond, A. and Sekhon, J. S. (2005). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies, *Review of Economics and Statistics* (0).

Doyle, M. W. and Sambanis, N. (2000). International peacebuilding: A theoretical and quantitative analysis, *American political science review* pp. 779–801.

Fortna, V. P. (2004). Does peacekeeping keep peace? international intervention and the duration of peace after civil war, *International Studies Quarterly* **48**(2): 269–292.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis* **20**(1): 25–46.

Hainmueller, J. and Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach, *Political Analysis* **22**(2): 143–168.

Iacus, S. M., King, G. and Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding, *Journal of the American Statistical Association* **106**(493): 345–361.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1): 243–263.

Kalton, G. (1983). Compensating for missing survey data.

King, G., Nielsen, R., Coberley, C., Pope, J. E. and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference, *Unpublished manuscript* **15**.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *The American Economic Review* pp. 604–620.

Lyall, J. (2010). Do democracies make inferior counterinsurgents? reassessing democracy's impact on war outcomes and duration, *International Organization* **64**(01): 167–192.

Micchelli, C. A., Xu, Y. and Zhang, H. (2006). Universal kernels, *The Journal of Machine Learning Research* **7**: 2651–2667.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**(2): 237–249.

Rubin, D. B. (1973). Matching to remove bias in observational studies, *Biometrics* pp. 159–183.

Rubin, D. B. (1990). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies, *Statistical Science* pp. 472–480.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials, *Statistics in medicine* **26**(1): 20–36.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, the MIT Press.

Sekhon, J. S. (2008). The neyman-rubin model of causal inference and estimation via matching methods, *The Oxford handbook of political methodology* pp. 271–299.

Smith, J. A. and Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods, *The American Economic Review* **91**(2): 112–118.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward, *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(1): 1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data, *Journal of the American Statistical Association* (just-accepted).

7. Appendix

7.1. Proof of Unbiasedness (Theorem 1)

Theorem 1 states that the weighted difference in means estimator using kernel balancing weights is unbiased for the sample average treatment effect on the treated (SATT) and the (population) ATT.

The SATT is similar to the ATT, but computes the average differences between the treatment and non-treatment potential outcome of the treated units actually sampled, rather than the expectation over the population distribution for the treated. The SATT is thus a more natural immediate target for an estimator.

$$SATT = \frac{1}{N_1} \sum_{i:D_i=1} Y_{1i} - \frac{1}{N_0} \sum_{i:D_i=0} Y_{0i} \tag{19}$$

Recall that the $\widehat{DIM}_w$ is defined as $\frac{1}{N_1} Y_{1i} - \sum_{D=0} w_i Y_{0i}$. Recall also that under the assumption $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top \theta$ (Assumption 2), $Y_{0i} = \phi(X_i)^\top \theta + \epsilon_i$ for $\mathbb{E}[\epsilon_i|X_i] = 0$.

Hence the error of the $\widehat{DIM}_w$ estimate for the SATT is then

$$\begin{aligned}
\widehat{DIM}_w - SATT &= \frac{1}{N_1} \sum_{i:D_i=1} Y_{0i} - \sum_{D_i=0} w_i Y_{0i} && (20)\\
&= \frac{1}{N_1} \sum_{i:D_i=1} \left(\phi(X_i)^\top \theta + \epsilon_i\right) - \sum_{i:D_i=0} w_i \left(\phi(X_i)^\top \theta + \epsilon_i\right) && (21)\\
&= \theta^\top \frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \theta^\top \sum_{i:D_i=0} w_i \phi(X_i) - \sum_{i:D_i=0} w_i \epsilon_i && (22)\\
&= \theta^\top \left(\frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) - \sum_{i:D_i=0} w_i \phi(X_i) \right) + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i && (23)\\
&= 0 + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i && (24)
\end{aligned}$$

The bias is the expectation of this quantity,

$$\begin{aligned}
bias &= \mathbb{E}\left[\widehat{DIM}_w - SATT\right] && (25)\\
&= \mathbb{E}\left[\frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i \right] = 0 && (26)
\end{aligned}$$

Remarks

Note that $\mathbb{E}[SATT] = ATT$, and so unbiasedness of $\widehat{DIM}_w$ for the SATT also implies unbiasedness for the ATT.

The assumption that $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top \theta$ is innocuous as $N \to \infty$, because the universal representation property of the Gaussian kernel ensures that the space of functions spanned by $\phi(X_i)^\top \theta$, which has representation $f(x_i) = \sum_j \alpha_j k(X_j, X_i)$, includes all continuous function. However, in

finite samples the quality of approximation is limited. Imagine the superposition of Gaussians view of this functions space: with too few observations, there are limits to the shapes that can be built by placing Gaussians at each observation and rescaling them. Even though highly non-linear, non-additive functions can still be well modeled with relatively small samples (see Hainmueller and Hazlett 2014), we may still wish to know how finite samples behave in terms of potential bias. Suppose that in truth, $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top\theta + h(X_i) + \epsilon_i$, where $h(X_i)$ is the misspecification error, an additive component that cannot be captured by $\phi(X_i)^\top\theta$ using the sample available and by definition orthogonal to the span of $\phi(X_i)$. In this case, the difference between $\widehat{DIM}_w$ and the SATT becomes

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i + \frac{1}{N_1}\sum_{i:D_i=1} h(X_i) - \sum_{i:D_i=0} w_i h(X_i) \tag{27}$$

Notice that bias due to misspecification occurs only if $h(X_i)$ has different means for the treated and controls (after weighting). That is, even if in a small sample $\mathbb{E}[Y_{0i}|X_i]$ cannot be well approximated, this is only problematic if the misspecification error, $h(X_i)$ is correlated with the treatment assignment after adjusting for differences on the other covariates through weighting. This is analogous to the biased caused by omitted variables in regression models.

7.2. Balance in $\mathbb{E}[\phi(X_i)]$ implies balance in $\mathbb{E}[Y_{0i}]$

The main text focuses principally on SATT estimation, and the implications of obtaining balance on $\phi(X_i)$ in the finite sample. However working with populations instead, we note that obtaining $\mathbb{E}[\phi(X_i)|D_i = 1] = \mathbb{E}_w[\phi(X_i)|D_i = 0]$ also implies $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_w[Y_{0i}|D_i = 0]$, where $\mathbb{E}_w[\cdot]$ designates an expectation taken over the w-weighted distribution of X:

$$\begin{aligned}
\mathbb{E}[Y_{0i}|D=1] &= \mathbb{E}_x\left[\mathbb{E}[Y_{0i}|X, D=1]\right] && (28)\\
&= \theta^\top \int \phi(x)p(x|D=1)dx && (29)\\
&= \theta^\top \mathbb{E}[\phi(x)|D=1] && (30)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_w[Y_{0i}|D=0] &= \mathbb{E}_{w,x}\left[\mathbb{E}[Y_{0i}|X, D=0]\right] && (31)\\
&= \theta^\top \int \phi(x)wp(x|D=0)dx && (32)\\
&= \theta^\top \mathbb{E}_w[\phi(x)|D=0] && (33)
\end{aligned}$$

Hence when balance of $\phi(X_i)$ for the treated and controls holds in expectations, we will have $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_w[Y_{0i}|D_i = 0]$, allowing a (weighted) difference in means to unbiasedly estimate the ATT.

7.3. Proof of proposition 1

Proposition 1 states: that for the mean row of $\mathbf{K}$ among the treated, $\overline{k_t} = \frac{1}{N_1}\mathbf{K_t}\mathbf{1}_{N_1}$ and the weighted mean row of $\mathbf{K}$ among the controls given by $\overline{k_c}(w) = \frac{\sum_i w_i k_i \mathbb{1}_{\{D_i=0\}}}{N_0}$, if $\overline{k_t} = \overline{k_c}(w)$, then $\overline{\phi_t} = \overline{\phi_c}$ where $\overline{\phi_t} = \frac{1}{N_1}\sum_{D_i=1}\phi(x_i)$ and $\overline{\phi_c} = \sum_{D_i=0}\phi(x_i)$.

This can be shown as follows.

$$\overline{k_T} = \sum_{i:D_i=0} w_i k_i \tag{34}$$

$$\frac{1}{N_1}\left[\sum_{i:D_i=1} k(X_i, X_1), \ldots, \sum_{i:D_i=1} k(X_i, X_N)\right] = \left[\sum_{i:D_i=0} w_i k(X_i, X_1), \ldots, \sum_{i:D_i=0} w_i k(X_i, X_N)\right] \tag{35}$$

$$\frac{1}{N_1}\sum_{i:D_i=1} [\langle\phi(X_i), \phi(X_1)\rangle, \cdots, \langle\phi(X_i), \phi(X_N)\rangle] = \sum_{i:D_i=0} w_i[\langle\phi(X_i), \phi(X_1)\rangle, \cdots, \langle\phi(X_i), \phi(X_N)\rangle] \tag{36}$$

$$\frac{1}{N_1}\sum_{i:D_i=1} \langle\phi(X_i), \phi(X_j)\rangle = \sum_{i:D_i=0} w_i\langle\phi(X_i), \phi(X_j)\rangle, \; \forall j \tag{37}$$

$$\langle\frac{1}{N_1}\sum_{i:D_i=1} \phi(X_i), \phi(X_j)\rangle = \langle\sum_{i:D_i=0} w_i\phi(X_i), \phi(X_j)\rangle \tag{38}$$

$$\langle\overline{\phi_t}, \phi(X_j)\rangle = \langle\sum_{i:D_i=0} w_i\phi(X_i), \phi(X_j)\rangle \tag{39}$$

$$\overline{\phi_t} = \sum_{i:D_i=0} w_i\phi(X_i) \tag{40}$$

7.3.1. Remarks

An intuitive interpretation of equation 38 is that each unit j is as close to the average treated unit as it is to the (weighted) average control unit, where distance is measured in the feature space $\phi(X)$. For the Gaussian kernel, $\langle\phi(X_i), \phi(X_j)\rangle$ is naturally interpretable as a similarity measure in the *input* space, since this quantity equals $k(X_j, X_i) = e^{-\frac{||X_j - X_i||^2}{s^2}}$. However, $\langle\phi(X_i), \phi(X_j)\rangle$ or $k(X_i, X_j)$ is more generally interpretable as similarity in the feature space as well. Note the squared Euclidean distance between two points X_i and X_j after mapping into $\phi(\cdot)$ is: $\|\phi(X_i) - \phi(X_j)\|^2 = \langle\phi(X_i) - \phi(X_j), \phi(X_i) - \phi(X_j)\rangle = \langle\phi(X_i), \phi(X_i)\rangle + \langle\phi(X_j), \phi(X_j)\rangle - 2\langle\phi(X_i), \phi(X_j)\rangle$. In the case of the Gaussian kernel, $\langle\phi(X_i), \phi(X_i)\rangle = 1$, so this distance reduces to $2(1 - \langle\phi(X_i), \phi(X_j)\rangle)$. In this sense, $\langle\phi(X_i), \phi(X_j)\rangle$ is as reasonable measure of similarity of position in the feature space, as it runs opposite to distance in this space.

Relatedly, a discriminant method of classifying observations as treated or control based on whether they are closer to the centroid of the treated or the centroid of the controls in $\phi(X)$ would be unable to classify any point.

7.4. Proof of proposition 2

Proposition 2 states that for a density estimator for the treated, $\hat{f}_{X|D=1}$, and for the (weighted) controls, $\hat{f}_{X|D=0,w}$, both constructed with kernel k with scale s^2, the choice of weights that ensures mean balance in the kernel matrix $\mathbf{K}$ also ensures $\hat{f}_{X|D=1} = \hat{f}_{X|D=0,w}$ at every location in $\mathcal{X}$ at which an observation is located.

As detailed in the main text, the expression $\frac{1}{N_1\sqrt{\pi s^2}} K_t \mathbf{1}_{N_1}$ places a multivariate standard normal density over each *treated* observation, sums these to construct a smooth density estimator at all points in $\mathcal{X}$, and evaluates the height of that joint density estimate at each of the points found

in the dataset. Likewise, $\frac{1}{N_0\sqrt{\pi s^2}}K_c\mathbf{1}_{N_0}$ estimates the density of the control units and returns its evaluated height at every datapoint in the dataset.

To reweight the controls would be to say that some units originally observed should be made more or less likely. This is achieved by changing the numerator of each weight $\frac{1}{N_0\sqrt{\pi s^2}}$ to some non-negative value other than 1. Letting the weights sum to 1 (rather than N_0), the reweighted density of the controls would be evaluated at each point in the dataset according to $\frac{1}{\sqrt{\pi s^2}}K_c w$, for vector of weights w. If weights are selected so that this equals the density of the treated:

$$\begin{aligned}
\frac{1}{N_1\sqrt{\pi s^2}}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \frac{1}{\sqrt{\pi s^2}}\mathbf{K_c}w \\
\frac{1}{N_1}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \mathbf{K_c}w \\
\overline{k_t} &= \mathbf{K_c}w \\
\overline{k_t} &= \overline{k_c(w)}
\end{aligned} \tag{41}$$

where the final line is the definition of mean balance in $\mathbf{K}$. Thus, the weights that achieve mean balance in $\mathbf{K}$ are precisely the right weights to achieve equivalence of the measured multivariate densities for the treated and controls at all points in the dataset.

7.4.1. Density Equalization Illustration

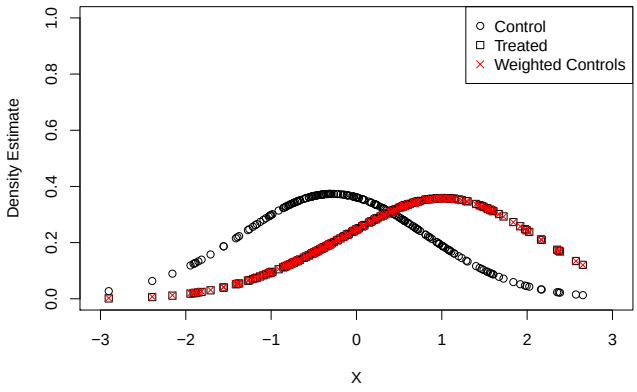
This example visualized the density estimates produced internally by kernel balancing using linear combinations of $\mathbf{K}$ as described above. Suppose X contains 200 observations from a standard normal distribution. Units are assigned to treatment with probability $1/(1+exp(2-2X))$, which produces approximately 2 control units for each treated unit. Figure 5 shows the resulting density plots, using density estimates provided by `kbal` in which the density of the treated is given by $\frac{1}{N_1\sqrt{\pi s^2}}\mathbf{K_t}\mathbf{1}_{N_1}$ and the density of the controls is given by $\frac{1}{N_0\sqrt{\pi s^2}}\mathbf{K_c}\mathbf{1}_{N_0}$. As shown, the density estimates for the treated at each observations X position (black squares) is initially very different from the density estimates for the controls taken at each observation (black circles). After weighting, however, the new density of the controls as measured at each observation (red `x`) matches that of the treated almost exactly.

Note that in multidimensional examples, the density becomes more difficult to visualize across each dimension, but it is still straightforward to compute and to think about the pointwise density estimates for the treated or control as measured at each observation's X value. In contrast to binning approaches such as CEM, equalizing density functions continuously in this way avoids difficult or arbitrary binning decisions, is tolerant of high dimensional data, and smoothly matches the densities in a continous fashion, resolving the within-bin discrepancies implied by CEM.

7.4.2. L_1, imbalance, and *numdims*

Recall that kernel balancing does not directly achieve mean balance on $\mathbf{K}$, but rather on the first *numdims* factors of $\mathbf{K}$ as determines by principal components analysis. This example examines the efficacy of this approach in minimizing the L_1 loss, and in minimizing imbalance on an unknown function of the data. Suppose we have 500 observations and 5 covariates, each with a standard normal distribution. Let $z = \sqrt{x_1^2 + x_2^2}$. This function impacts treatment assignment, with the

Figure 5: Density-Equalizing Property of Kernel Balancing



Plot showing the density-equalization property of kernel balancing. For 200 observations of $X \sim N(0, 1)$, treatment is assigned according to $Pr(treatment) = 1/(1 + exp(2 - 2X))$, producing approximately two control units for each treated unit. Black squares indicate the density of the treated, as evaluated at each observation's location in the dataset (and given the choice of kernel and s^2). Black circles indicate the density of (unweighted) controls. The treated and control are seen to be drawn from different distributions, owing to the treatment assignment process. Red x's show the new density of the controls, after weighting by `kbal`. The reweighted density is nearly indistinguishable from the density of the treated, owing to the density equalization property of kernel balancing.

probability of treatment being given by $logit^{-1}(z - 2)$, which produces approximately two control units for each treated unit.

In figure 6, the value of *numdims* – the number of factors of **K** retained for purposes of balancing – is increased from a minimum of 2 up to 100. As expected, both L_1 and the mean imbalance on z taken after weighting improve as *numdims* is first increased, and then worsen beyond some choice of *numdims*. Most importantly, while the balance on z is unobservable in the case of unknown confounders, L_1 is observable, and improvements in L_1 track very closely to improvements in the balance of z. Accordingly, selecting *numdims* to minimize L_1 appears to be a viable strategy for selecting the value that also minimizes imbalance on unseen functions of the data.

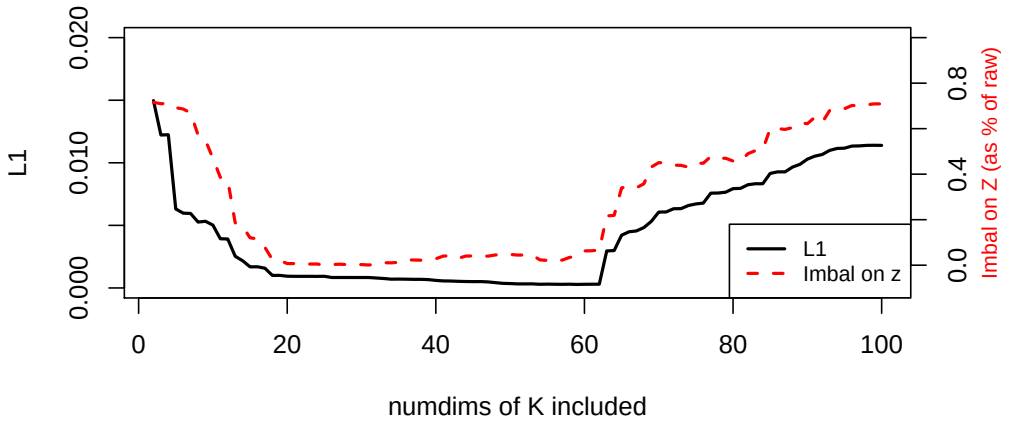
7.5. Optional Trimming of the Treated

In some cases, balance can be greatly improved with less variable (and thus more efficient) weights if the most difficult-to-match treated units are trimmed. In estimating an ATT, control units in areas with very low density of treated units can always be down-weighted (or dropped if the weight goes to zero), but treated units in areas unpopulated by control units pose a greater problem. These areas may prevent any suitable weighting solution, or may place extremely large (and thus ineffecient) weights on a small set of controls.

While estimates drawn from samples in which the treated are trimmed no longer represent the ATT with respect to the original population, they can be considered a local or sample average treatment effect within the remaining population. King et al. (2011) refer similarly to a "feasible sample average treatment effect on the treated" (FSATT), based on only the treated units for which sufficiently close matches can be found. In any case, the discarded units can be characterized to learn how the inferential population has changed.

However, even when the investigator is willing to change the population of interest by trimming the treated, it is not always clear on what basis trimming should be done. In kernel balancing, trimming of the treated can be (optionally) employed by using the multivariate density interpreta-

Figure 6: L_1 distance and imbalance on an unknown confounder, by *numdims*



This example shows the relationship between the number of components of **K** that get balanced upon (*numdims*), the multivariate imbalance (L_1), and balance on confounder z. L_1 generally improves as *numdims* is increased at first, but beyond approximately 50 dimensions, numerical instability produces less desirable results and a higher L_1 imbalance. While the confounder represented by z in this case would generally be unobservable, balance on z is optimized where L_1 finds its minimum, which is observable.

tion given above. Specifically, the density estimators at all points is constructed using the kernel matrix. Then, treated units are trimmed if $\frac{f_{X|D=1}(x_i)}{f_{X|D=0}(x_i)}$ exceeds the parameter *trimratio*. The value of *trimratio* can be set by the investigator based on qualitative considerations, inspection of the typical ratio of densities, a willingness to trim up to a certain percent of the sample, or performance on L_1. Whatever approach is taken to determine a suitable level of *trimratio*, `kbal` produces a list of the trimmed units, which the investigator can examine to determine how the inferential population has changed.

7.6. Additional Example: Are Democracies Inferior Counterinsurgents?

Decades of research in international relations has argued that democracies are poor counterinsurgents (see Lyall 2010 for a review). Democracies, as the argument goes, are (1) sensitive to public backlash against wars that get more costly in blood or treasure than originally expected, (2) are unable to control the media in order to supress this backlash, and (3) often respect international prohibitions on brutal tactics that may be needed to obtain a quick victory. Each of these makes them more prone to withdrawal from countinsurgency operations, which often become long and bloody wars of attrition. Empirical work on this question was significantly advanced by Lyall (2010), who points out that previous work (1) often examined only democracies rather, than a universe of cases with variation on polity type, and (2) did little to overcome the non-random assignment of democracy, and particular, the selection effects by which democracies may choose to fight different types of counterinsurgencies than non-democracies.

Lyall (2010) overcomes these shortcomings by constructing a dataset covering the period of 1800-2005, in which the polity type of the countinsurgent regimes vary. Matching is then used to adjust for observable differences between the conflicts selected by democracies and non-democracies, using one-to-one nearest neighbor matching on a series of covariates. These covariates are: a dummy for

whether the counterinsurgent is an occupier (*occupier*), a measure of support and sanctuary for insurgents from neighboring countries (*support*), a measure of state power (*power*), mechanization of the military (*mechanized*), *elevation*, *distance* from the state capital to the war zone, a dummy for whether a state is in the first two years of independence (*new state*), a *cold war* dummy, the number of *languages* spoken in the country, and the *year* in which the conflict began.

In a battery of analyses with varying modeling approaches, Lyall (2010) finds that democracy, measured as a polity score of at least 7 in the specifications replicated here, has no relationship to success or failure in counter insurgency, either in the raw data or in the matched sample.

While the credibility of this estimate as a causal quantity depends on the absence of unobserved confounders, we can nevertheless assess whether the procedures used to adjust for observed covariates were sufficient, or whether an inability to achieve mean balance on some functions of the covariates may have led to bias even in the absence of unobserved confounders.

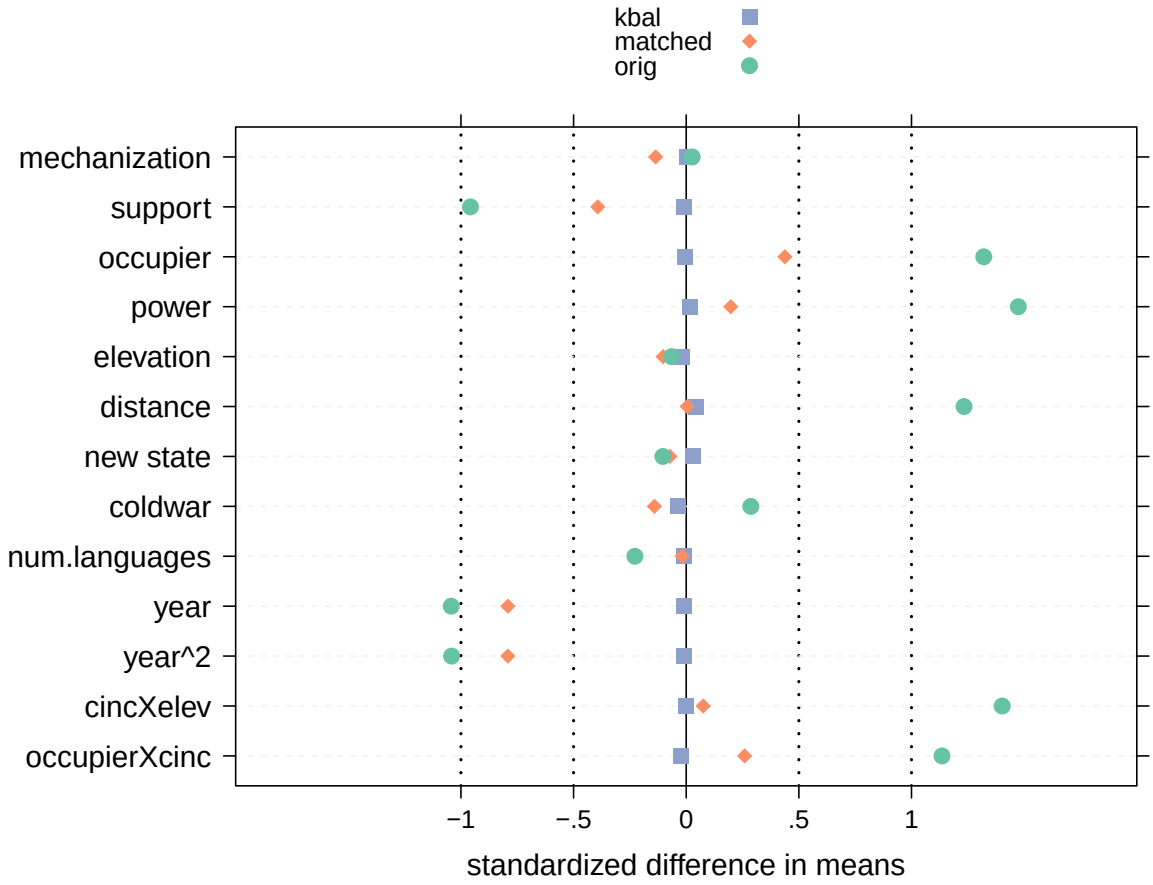
Here I reexamine these findings using the post-1945 portion of the data, which includes 35 counterinsurgencies by democracies and 100 by non-democracies, and is used in many of the analyses in Lyall (2010). The 1945 period is the only one with complete data on the covariates used for balancing here, but is also the period in which the logic of democratic vulnerability is expected to be most relevant.

First, I assess balance. As shown in figure 7, numerous covariates are badly imbalanced in the original dataset (circles), where imbalance is measured on the x-axis by the standardized difference in means. This balance improves somewhat under matching (diamonds), but improves far more under kernel balancing (squares). Note that imbalance is shown both on the variables used in the matching/weighting algorithms (the first ten covariates up to and including *year*), as well as several others that were not explicitly included in the balancing procedure: $year^2$, and two multiplicative interactions that were particularly predicted of treatment status in the original data. Kernel balancing produces good balance on both the included covariates, and functions of them.

Next, I use the matched and weighted data to estimate the effect of democracy on counterinsurgency success. For this, I simply use linear probability models (LPM) to regress a dummy for victory (1) or defeat (0) on covariates according to five different specifications. While Lyall (2010) used a number of other approaches, including logistic regression, some of these models suffer "separation" under the specifications attempted here. This causes observations and variables to effectively drop out of the analysis, producing variability in effect estimates that are due only to this artefact of logistic regression and not due to any meaningful change in the relationship among the variables. Linear models do not suffer this problem, and provide a well defined approximation to the conditional expectation function, allowing valid estimation of the changing probability of victory associated with changes in the treatement variable, *democracy*. The first three specifications used are (1) *raw* regresses the outcome directly on *democracy* without covariates (and is equivalent to difference-in-means);(2) *orig* uses the same covariates as Lyall (2010), which are all those variables balanced on except for *year*, (3) *time* reincludes *year* as well as $year^2$ to flexibly model the effects of time. The final two models, *occupier1* (4) and *occupier2* (5), add flexibility by including interactions of *occupier* with other variables in the model. These interactions were chosen because analysis with KRLS revealed that interactions with *occupier* were particularly predictive of the outcome.

Figure 8 shows results for the matched and kernel balanced samples with 95% confidence intervals. Under matching, the effect varies considerably depending on the choice of model. No estimate

Figure 7: Balance: Democracies vs. Non-democracies and the Counterinsurgencies they Fight



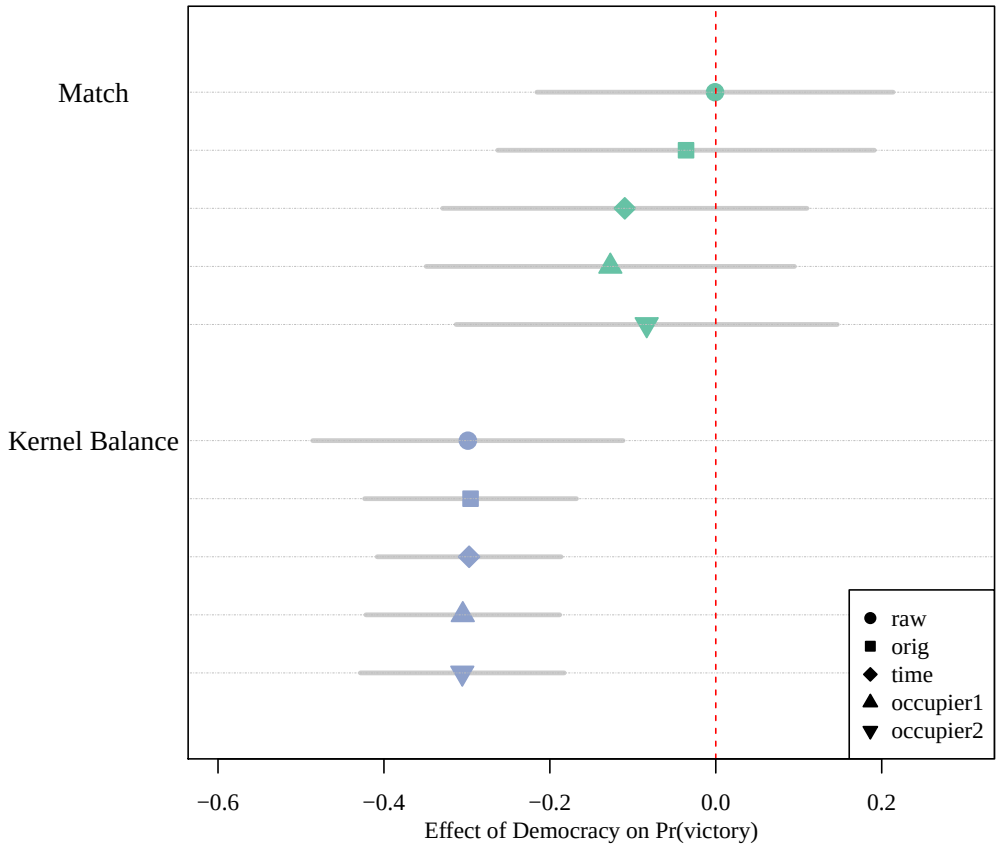
Balance in post-1945 sample of Lyall (2010). Imbalance, measured as the difference in means divided by the standard deviation, is shown on the x- axis. Democracies (treated) and non-democracies (controls) vary widely on numerous covariates. The matched sample (diamonds) shows somewhat improved balance over the original sample, but imbalances remain on numerous characteristics. Balance is considerably improved by kernel balancing (squares). The rows at or above *year* show imbalance on characteristics explicitly included in the balancing procedures. Those below *year* show imbalance on characteristics not explicitly included.

is significantly different from zero, however. In stark contrast, kernel balancing producing estimates that are essentially invariant to the choice of model. Each kernel balancing estimate is between -0.26 and -0.27, indicating that democracy is associated with a 26 to 27 percentage point lower probability of success in fighting counterinsurgencies. This is a very large effect, both statistically and substantively, given that the overall success rate is only 33% in the post-1945 sample.

7.7. Are democracies more selective?

One puzzle regarding the claim that democracies are inferior counterinsurgents has been why democracies, whatever their weaknesses as counerinsurgents, are not also better able to "select into" conflicts they are more likely to win. The same qualities that are theorized to make democracies more susceptible to defeat against insurgents – public accountability and media freedoms – might also push democracies to more carefully select what counterinsurgency operations they engage in.

Figure 8: Effect of Democracy on Counterinsurgency Success



Effect of democracy on counterinsurgency success in post-1945 sample of Lyall (2010) using matching or kernel balancing for pre-processing followed by five different estimation procedures. Under matching, effect estimates remain highly variable, but none are significantly difference from zero. Kernel balancing shows remarkably stable estimates over the five estimation procedures, even when no covariates are included (*raw*). Results from kernel balancing are consistently in the -0.26 to -0.27 range and significantly different from zero, indicating that democracy is associated with a substantively large deficit in the ability to win counterinsurgencies.

The findings suggest that such a selection may occur. Specifically, the naive effect estimate obtained by a simple difference in mean probability of victory (on the unweighted sample) is -0.10 ($p = 0.13$). Recall that this difference in means can be decomposed,

$$\begin{aligned}\mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=0] &= \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1] + \mathbb{E}[Y(0)|D=1] - \mathbb{E}[Y(0)|D=0] \\ &= ATT + [\mathbb{E}[Y(0)|D=1] - \mathbb{E}[Y(0)|D=0]]\end{aligned}$$

That is, the naive difference in means is the average treatment effect on the treated (had they fought in the same types of cases), plus a selection effect indicating how democracies and non-democracies differ in their probabilities of victory based only on fighting different types of cases (i.e. in the absence of any effect of democracy). Since we know the ATT estimate and the raw difference in means, we can estimate the selection effect to be about 17 percentage points more

likely to end in victory. While simple, this decomposition suggests that democracies do choose counterinsurgencies somewhat "wisely", but are also less likely to win a given a counterinsurgency once this selection is accounted for.