Statistica Sinica Preprint No: SS-2017-0555	
Title	Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects
Manuscript ID	SS-2017-0555
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0555
Complete List of Authors	Chad Hazlett
Corresponding Author	Chad Hazlett
E-mail	chazlett@ucla.edu
Notice: Accepted version subject to English editing.	

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

Chad Hazlett
Departments of Statistics & Political Science,
University of California Los Angeles*

July 18, 2018

Abstract

Matching and weighting methods are widely used to estimate causal effects when adjusting for a set of observables is required. Matching is appealing for its non-parametric nature, but with continuous variables, is not guaranteed to remove bias. Weighting techniques choose weights on units to ensure pre-specified functions of the covariates have equal (weighted) means for the treated and control group. This assures unbiased effect estimation only when the potential outcomes are linear in those pre-specified functions of the observables. Kernel balancing begins by assuming the expectation of the non-treatment potential outcome conditional on the covariates falls in a large, flexible space of functions associated with a kernel. It then constructs linear bases for this function space and achieves approximate balance on these bases. A worst-case bound on the bias due to this approximation is given and is the target of minimization. Relative to current practice, kernel balancing offers one reasoned solution to the long-standing question of which functions of the covariates investigators should attempt to achieve (and check) balance on. Further, these weights are also those that would make the estimated multivariate density of covariates approximately the same for the treated and control groups, when the same choice of kernel is used to estimate those densities. The approach is fully automated up to the choice of a kernel and smoothing parameter, for which default options and guidelines are provided. An `R` package, `KBAL`, implements this approach.

Keywords: causal inference, statistical learning, covariate balance, weighting, matching

*I thank Jens Hainmueller, Teppei Yamamoto, Brandon Stewart, Kosuke Imai, Mark Ratkovic, Jeff Lewis, Mark Handcock, and Arash Amini for valuable feedback and support on this project.

1 Introduction

It is often necessary to adjust for covariates in seeking to make causal inferences from observational data under the assumption of no unobserved confounding or conditional ignorability. Matching and weighting techniques seek to adjust for covariates, making the distribution of these covariates similar in the treated and control groups. However, when exact matching is not possible – such as when continuous variables are included – these methods can fail to implement the conditioning or adjustment for which they are intended. For concreteness, suppose an investigator matches or weights on continuous, pre-treatment covariates X_1 and X_2, but it is the ratio, X_1/X_2, that is critical. Specifically, suppose that both the potential outcomes and the probability of taking the treatment are monotonically increasing in X_1/X_2. Though matching has a desirable non-parametric nature, the failure to find exact matches with multiple continuous variables is problematic in finite samples: among treated and control units matched to each other, the treated unit will on average be higher on X_1/X_2 than the control unit, and thus higher in its expected potential outcomes than the controls unit. This "matching discrepancy" does not dissipate quickly with increasing sample size, and previous work such as Abadie and Imbens (2006) has shown the resulting bias and lack of $\sqrt{N}$ consistency of matching estimators for this reason.

By comparison, standard weighting approaches can (when feasible) obtain exact or approximate balance on desired moments, such as means of X_1 and X_2, and so may seem to offer a technology for adjustment that sidesteps the matching discrepancy problem. However, this sacrifices the non-parametric quality of matching. If a weighting estimator obtains equal means for the treated and control groups on both X_1 and X_2 (referred to here as "mean balance"), this does not in general imply that X_1/X_2 or other non-linear functions of X_1 and X_2 have equal means for the two groups. The difference between these groups on the outcome is thus contaminated by differences on important functions of observables, resulting in bias. In short, both weighting and matching as described thus far would fail to make the treated and control groups "comparable" on the observables. The desired adjustment for observables required to estimate effects is thus simply not performed.

Such biases could be avoided if the investigator happened to know that X_1/X_2 was the critical function of the observables to match or weight on. However, rarely can we expect investigators to have sufficient theoretical knowledge to unfailingly guess these functional forms. Moreover, allowing the

investigator the freedom to guess at such functional forms creates opportunities for selective reporting of results. The simulations in Section 3.1 further examine this hypothetical example, showing how simple non-linear functions of the observables can generate large biases when using state-of-the-art matching and weighting estimators, even when bias adjustment procedures (Abadie and Imbens, 2011) are applied. Kernel balancing mitigates this problem, achieving nearly equal means on X_1/X_2 without the investigator knowing of it's importance. This robustness, however, comes at the cost of a mild assumption on the form of the (non-treatment) potential outcome surface, $\mathbb{E}[Y_{0i}|X_i]$.

Delaying technical details, the fundamental idea behind kernel balancing is straightforward. First it should be said that, as with regression, matching, sub-classification, and other covariate adjustment procedures, we assume that the set of variables one should adjust for to achieve causal identification are indeed observed.[1] Next, we assume the regression surface for the non-treatment potential outcome (Y_{0i}) conditional on the adjustment covariates falls in the (Reproducing Kernel Hilbert) space associated with a choice of kernel. Here, I propose using a Gaussian kernel, as the corresponding function space is suitable to a wide variety of smoothly varying outcomes. The practical meaning of this assumption and an interpretation of the resulting function space is provided herein. Third, the empirical kernel matrix $\mathbf{K}$ with rows K_i forms a basis set for the regression function, $\mathbb{E}[Y_{0i}|X_i]$. This amounts simply to a change of bases, from $X_i \in \mathcal{R}^P$ to $K_i \in \mathcal{R}^N$, enabling us to access highly flexible and complex functions rather than those simply linear in X_i. Fourth, having chosen these bases, kernel balancing finds weights on the control units such that the weighted average K_i among the control units approximately equals the (unweighted) average K_i among the treated units. This is the key step: since the regression surface for Y_{0i} is linear in K_i, achieving (approximately) equal means on K_i ensures (approximately) equal means on Y_{0i} for the treated and weighted control groups, without having to fit any model. The worst-case bound on the remaining bias is minimized when choosing the weights. Finally, a simple difference in (weighted) means can then be used to estimate the average treatment effect on the treated (ATT). The remainder of the paper expands upon this logic.

[1]Throughout, we assume that the investigator has correctly chosen the set of covariates that must be conditioned on and are not conditioning on covariates that would only increase bias, such as post-treatment variables or colliders (Pearl, 2009). Once the set $\{X\}$ of conditioning variables has been chosen to satisfy the "adjustment criteria" in a graphical causal model, it can also be said that conditional ignorability holds given $\{X\}$, i.e. $Y(d) \perp\!\!\!\perp D|X$ (Elwert, 2013). This conditional ignorability is the central identification assumption used throughout.

In principal if one could obtain equal multivariate densities of covariates for the treated and control group, this would non-parametrically and fully adjust for covariates: in the absence of confounders, this would ensure $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$, which ensures unbiasedness of the difference in means for the ATT, *regardless of the form of* $\mathbb{E}[Y_{0i}|X_i]$. The difficulties with such a "full multivariate density equality" approach are practical. Such equality cannot be verified in full, except in cases where there is a small number of discrete covariates with small numbers of categories each. The simple alternative approach taken here is to first assume that $\mathbb{E}[Y_{0i}|X_i]$ is linear in some set of bases, and attempt to achieve equal means on these. Any matching or weighting estimator that can achieve mean balance on some $\phi(X_i)$ could be justified, if $\mathbb{E}[Y_{0i}|X_i]$ is indeed linear in $\phi(X_i)$. The analytical framework for kernel balancing elaborates upon this idea and proposes a set of choices for implementing it: if asserting balance on $\phi(X_i)$ guarantees unbiasedness only when $\mathbb{E}[Y_{0i}|X_i]$ is linear in $\phi(X_i)$, let us chose a basis expansion $\phi(\cdot)$ that is very general and likely to contain a close approximation to $\mathbb{E}[Y_{0i}|X_i]$. Kernels provide one convenient and powerful choice for this purpose. This approach makes explicit what balance tests would ideally be satisfied to achieve zero bias.

Further, while full multivariate density equality is not the aim of kernel balancing, an illuminating connection emerges between that approach and the "balance on kernel-derive bases" approach: the weights chosen by kernel balancing to achieve mean balance on the chosen bases are exactly those that would equalize the *estimated* empirical multivariate distributions of covariates for the treated and control, *when the same kernel is used for density estimation.* This reveals a direct link between (1) the assumption one is willing to make about the space of outcome models, and (2) the choice of a smoother such that the smoothed multivariate distribution of covariates are made equal in the treated and control groups.

To briefly place kernel balancing in context, compared to approaches that depend on fitting outcome models (such as regression), kernel balancing still relies on an assumed outcome model space, though no outcome model ever need be fitted. By contrast, like matching, dependence on such an outcome model is reduced, because the (weighted) densities of treated and control are made similar to allow for comparison of the two samples on their outcomes, rather than relying on strong modeling assumptions to bridge potentially large gaps between the location of control and treated observations. On the

other hand, kernel balancing also differs from existing matching and weighting approaches. Even when matching methods achieve perfect balance according to whichever imbalance measures they employ, these balance metrics typically check only for equal means on the covariates, or other moments as specified by the user. Unfortunately, as in the brief example above, matching discrepancies that occur under inexact matching can give rise to imbalances on unchecked functions of the covariates, which in turn can generate biased ATT estimates. Though debiasing methods have been proposed (Abadie and Imbens, 2011), they work by turning to a functional form assumption and thus are not always effective, as illustrated below. Turning to propensity score approaches (Rosenbaum and Rubin, 1983), kernel balancing differs in that like matching and weighting estimators described above, it requires no functional form assumption for the probability of receiving treatment given the covariates. This avoids the severe biases that can occur due to possible misspecification of the propensity score (see e.g., Smith and Todd, 2005; Kang and Schafer, 2007). Finally, the method is most similar to other weighting or calibration procedures that do not model treatment assignment, instead obtaining balance on covariates (such as Hainmueller, 2012; Zubizarreta, 2015), not to mention analogous survey weighting procedures going back at least to Deming and Stephan (1940). One contribution of kernel balancing relative to these procedures is that it makes explicit, then weakens, the linear functional form assumption inherent in these weighting and calibration approaches. From the investigator's perspective, the most immediate and practical contribution of kernel balancing is that it provides practioners with one principled and automated answer to the question of what functions of the covariates should be made to have approximately equal means, given an assumption that the outcome lies in a flexible, smooth space of models. Outside the causal inference framework, the same procedure can be used to reweight survey data to match a population of interest, not only on the means of the covariates but on a large space of smooth functions of those covariates.

In what follows, Section 2 provides the analytical framework and develops the method. Section 3 provides a basic simulation, highlighting the dangers inherent in other methods under reasonable conditions and demonstrating kernel balancing as a potential solution. In Section 4, I provide an empirical demonstration of the method's effectiveness in recovering an experimental benchmark from observational data, using the National Supported Work demontration (LaLonde, 1986). Section 5

discusses the implications of this procedure, additional details, and further comparisons to existing matching, weighting, regression, and propensity score approaches. Section 6 concludes. Additional remarks, guidelines, proofs, and empirical examples can be found in the the Online Appendix.

2 Framework for Kernel Balancing

2.1 Notation

This section sets up the problem of ATT estimation, then describes the main ideas of the kernel balancing approach as a method to achieve covariate adjustment.

We first set notation. Using the Neyman-Rubin potential outcomes framework (see e.g. Rubin, 1990; Splawa-Neyman et al., 1990) 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, taking values $x \in \mathcal{X}$ where support $\mathcal{X}$ lies in $\mathbb{R}^P$. For all i, assume that draws of the random variables $\{Y_{1i}, Y_{0i}, X_i, D_i\}$ are taken independently from common joint density $p(Y_1, Y_0, X, D)$. The set of covariates in $\{X\}$ is assumed here to be the set that the investigator must condition upon in order to achieve a causal estimate by ensuring that treatment assignment is ignorable with respect to the potential outcomes, conditionally on the covariates, as assumed under "conditional ignorability",

Assumption 1 (Conditional Ignorability) *The potential outcomes are conditionally ignorable if*

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i \mid X_i$$

where Y_{0i} and Y_{1i} are the non-treatment and treatment potential outcomes, D_i is treatment status, and X_i is a vector of observed, pre-treatment covariates.

Kernel balancing chooses a space of models relating the (potential) outcomes to the covariates, which determines what functions of the covariates must have equal expectations in the treatment and control groups to ensure the non-treatment potential outcomes have equal expectations. Thus, rather

than fitting a propensity score model, we impose assumptions on the ways in which the covariates relate to potential outcomes. Specifically, assume $X \in \mathbb{R}^P$ is a set of covariates or characteristics satisfying Assumption 1, and $\phi(X) : \mathbb{R}^P \mapsto \mathbb{R}^Q$, where Q may be (much) larger than P (or N), giving an expanded set of characteristics or features to be used as a set of basis functions.[2] The specific nature of $\phi(\cdot)$ used in kernel balancing will relate to a choice of kernel (with a Gaussian kernel used in the particular implementation here, described further below). For the moment, the key feature of $\phi(\cdot)$ needed is that it is a sufficiently rich, non-linear expansion such that $\mathbb{E}[Y_{0i}|X_i = x]$ can be well fitted as a linear function of $\phi(x)$:[3]

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. there exists $\theta \in \mathbb{R}^Q$ and $\phi(\cdot) : \mathbb{R}^P \mapsto \mathbb{R}^Q$ such that*

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

We will soon see that the choice of $\phi(X_i)$ to be used will be a very general one associated with a kernel, with special attention to the case of the Gaussian kernel. This will allow the function space $\phi(X_i)^\top\theta$ to capture all continuous functions as $N \to \infty$. More importantly, in finite samples, this space can be understood as the smooth and flexible space of functions that can be built by placing (Gaussian) kernels over the observations, rescaling them as needed, and summing them. This is described at length below and particularly in Section 5.3. In addition, potential violations of Assumption 2 bias the resulting estimate only to the degree that components of $\mathbb{E}[Y_{0i}|X_i]$ not in the span on $\phi(X_i)$ are correlated with treatment D_i (see Appendix A.2.1).

[2]Two details are worth noting regarding Assumption 1. First, for purposes of ATT estimation alone, it could be weakened to $Y_{0i} \perp\!\!\!\perp D_i \mid X_i$. This is effectively because the Y_{1i} values needed in ATT estimation are observed; assumptions need not be made about how they can be proxied by other values. Second, the conditional ignorability argument is usually paired with a "positivity" or "common support" assumption requiring that $0 < Pr(D_i|X_i) < 1, \quad \forall\ X_i \in \mathcal{X}$. Such a requirement is especially evident for propensity-score based estimators. However it is not required here, as a consequence of having made an assumption about the regression surface of Y_{0i} in terms of basis functions (see Assumption 2).

[3]Note that a similar assumption can be made regarding $\mathbb{E}[Y_{1i}|X_i]$, and is required for analysis of the average treatment effect (ATE) or average treatment effect on controls (ATC). This paper focuses first on the ATT for ease of exposition.

2.2 Population ATT and DIM

Let us take the population ATT, $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$, as our quantity of interest. This can now be expressed as

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \int \mathbb{E}[Y_{0i}|x, D_i = 1]p(x|D_i = 1)dx \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \int \phi(x)^\top \theta p(x|D_i = 1)dx \end{aligned}$$

where $\mathbb{E}[Y_{0i}|x, D_i = 1] = \mathbb{E}[Y_{0i}|x]$ due to Assumption 1, and $p(x|D_i = 1)$ is the density of X_i conditional on $D_i = 1$. We will examine the (population) difference in means estimator (DIM), and seek to determine the conditions under which it is equal to the ATT. The DIM is given by

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

which replaces the unobservable second term in the ATT expression ($\mathbb{E}[Y_{0i}|D_i = 1]$) with its identifiable counterpart, $\mathbb{E}[Y_{0i}|D_i = 0]$. Rewriting this term using Assumption 2,

$$\begin{aligned} DIM &= \mathbb{E}[Y_{1i}|D_i = 1] - \int \mathbb{E}[Y_{0i}|x, D_i = 0]p(x|D_i = 0)dx \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \int \phi(x)^\top \theta p(x|D_i = 0)dx \end{aligned}$$

We can now see that without further adjustment, the DIM would equal the ATT only when

$$\int \phi(x)^\top \theta p(x|D_i = 0)dx = \int \phi(x)^\top \theta p(x|D_i = 1)dx \tag{2}$$

which holds for any θ when:

$$\int \phi(x)p(x|D_i = 0)dx = \int \phi(x)p(x|D_i = 1)dx$$
$$\mathbb{E}[\phi(X_i)|D_i = 0] = \mathbb{E}[\phi(X_i)|D_i = 1] \tag{3}$$

Both Expressions 2 and 3 have very useful direct interpretations. The equality in (2) can be interpreted as requiring $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$ in order for the population DIM to be the same as the ATT. Indeed, such mean independence could be used in place of the stronger full independence assumption (Assumption 1) when average treatment effects are all that are required.

However, it is Equation 3 that suggests a natural strategy for estimation: due to the linearity of the assumed function space for $\mathbb{E}[Y_{0i}|X_i]$ (Assumption 2), we obtain $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$ whenever $\mathbb{E}[\phi(X_i)|D_i = 0] = \mathbb{E}[\phi(X_i)|D_i = 1]$, regardless of θ – and without need of estimating it. We exploit this fact in the next section.

2.3 Achieving mean balance on $\phi(X_i)$ by weighting: Ideal case

Having chosen suitable $\phi(\cdot)$ as bases, we need only choose weights that produce equal means on $\phi(X_i)$ for the controls group and treated group, rather than on the original X_i. Let us refer to this as "mean balance on $\phi(X_i)$." Under Assumptions 1 and 2, achieving exact mean balance on $\phi(X_i)$ would ensure equal $\mathbb{E}[Y_{0i}]$ for the (weighted) controls and treated groups – which in turn would allow ATT estimation by a weighted difference in means estimators described below. Unfortunately, exact mean balance on $\phi(X_i)$ will be infeasible when $\phi(\cdot)$ is high dimensional, as proposed here. Instead, a feasible, approximate balancing approach is employed. We will then bound the bias that could occur due to this approximate rather than perfect balance, and examine the practical merit of this procedure through both simulations and applied examples.

Consider an adjustment procedure involving a function of the covariates $\tilde{g}(X_i)$, with the property that:

$$\begin{aligned}
\int \phi(x)^\top \theta \tilde{g}(x) p(x|D_i = 0)dx &= \int \phi(x)^\top \theta p(x|D_i = 1)dx \\
\int \phi(x)[\tilde{g}(x) p(x|D_i = 0)]dx &= \int \phi(x) p(x|D_i = 1)dx \\
\int \phi(x) g(x)dx &= \int \phi(x) p(x|D_i = 1)dx \\
\mathbb{E}_g[\phi(X_i)|D_i = 0] &= \mathbb{E}[\phi(X_i)|D_i = 1] \qquad (4)
\end{aligned}$$

where $g(x) = \tilde{g}(x)p(x|D_i = 0)$ is scaled so that $\int g(x)d(x) = 1$. This effectively gives us a new density, which we integrate over to obtain a "g-weighted" expectation of $\phi(X_i)$ among the controls. Setting $\tilde{g}(x) = \frac{p(x|D_i=1)}{p(x|D_i=0)}$ is one natural choice that satisfies this, directly making $g(x) = p(x|D_i = 1)$ (see Section 5 for equivalence to inverse propensity score weighting). However, any choice $g(x)$ satisfying Equation 4 makes the expectation of $\phi(X_i)$ the same for the treated and control.

Putting these pieces together, the ATT is identified by a DIM estimator modified by these weights,

$$DIM_w = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}_g[Y_{0i}|D_i = 0] \qquad (5)$$

To review thus far, we have simply established that: in the absence of unobserved confounders (Assumption 1) and linearity of the conditional expectation of Y_{0i} in $\phi(X_i)$ (Assumption 2), the DIM equals the ATT in the population when a $g(x)$ can be found such that the g-weighted expectation of $\phi(X_i)$ among the controls equals that unweighted expectation of $\phi(X_i)$ among the treated. In short, we have chosen bases for the expected non-treatment potential outcome, and ensured equal expectations on each of these bases, in turn ensuring equal expected non-treatment potential outcomes for the treated and control group. This, if achievable, would make the expectation of the non-treatment potential outcome for the untreated equal to that of the treated, i.e. $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_g[Y_{0i}|D_i = 0]$, in turn ensuring that the weighted DIM equals the ATT.

2.4 Sample DIM and Weights

We now turn to the sample and corresponding choice of weights for a plug-in estimator. Let N_0 equal the number of control observations and N_1 the number of treated observations. We estimate $\mathbb{E}[\phi(X_i)|D_i = 1]$ in Equation 4 with its sample analog, $\frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i)$. For the g-weighted expected non-treatment outcome among the controls, we also replace the expectation with the sample mean, and the "g-weights" with finite sample weights $w_1, ..., w_{N_0}$ that solve the sample moment constraints, where all $w_i \geq 0$ and $\sum_i w_i = 1$. Altogether then the sample conditions are given by

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

subject to the conditions $w_i \geq 0$, $\sum w_i = 1$. With w chosen this way (see Section 2.8) we can construct the sample estimator for the DIM, $\widehat{DIM}$. Working from Equation 5, we replace each expectation in the DIM with the corresponding empirical mean to define our estimator,

$$\widehat{DIM}_w = \frac{1}{N_1}\sum_{i:D_i=1} Y_i - \sum_{i:D_i=0} w_i Y_i \tag{6}$$

This brings us to the main result under exact mean balance,

Theorem 1 (Unbiasedness of Weighted Difference in Means for the ATT) *Consider the weighted difference in means estimator,*

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

$$\textit{where } w \textit{ satisfies } \sum_{i:D_i=0}\phi(X_i)w_i = \frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i), \textit{ subject to } \sum_i w_i = 1 \textit{ and } w_i > 0, \forall i$$

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 ATT, taken over common joint density* $p(X, Y_1, Y_0, D)$.

An alternative derivation begins with a given sample, showing that this weighted difference in means is unbiased for the sample average treatment effect (SATT), which in turn is unbiased for the population ATT under random sampling (See Appendix A.2). That approach also leads to an analysis of the finite sample bias under failure of Assumption 2, i.e. when $\mathbb{E}[Y_{0i}|X_i]$ is not fully linear in $\phi(X)$. The result indicates that bias is introduced only when the component of the regression surface ($\mathbb{E}[Y_{0i}|X_i]$) not linear in $\phi(X)$ is correlated with treatment assignment (Appendix A.2.1).

2.5 Kernels based construction of $\phi(\cdot)$

A wide range of basis expansions $\phi(\cdot)$ could in principle be chosen under this estimation framework. Here, the proposal is not to chose $\phi(\cdot)$ directly, but implicitly through a choice of kernel, which will generate an N-dimensional vector of features on which equal means can instead be achieved.

2.5.1 Kernel Notation

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. While numerous kernels could be used in this procedure, for reasons discussed below we continue here specifically with the Gaussian kernel:

$$k(X_j, X_i) = e^{-\frac{||X_j - X_i||^2}{b}} \tag{7}$$

Note that $k(X_i, X_j)$ produces values between 0 and 1 interpretable 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 b might be called "scale", because it governs how close X_i and X_j must be in a Euclidean sense to be deemed similar (see A.10 regarding the choice of b). It is common to rescale each covariate to have variance 1 prior to computing $k(X_i, X_j)$. This ensures results will be invariant to unit-of-measure decisions. Let the symmetric, matrix $\mathbf{K}$ be the the N-by-N positive semi-definite (PSD) kernel matrix, with elements $\mathbf{K}_{i,j} = k(X_i, X_j)$. Finally, let the i^{th} row (or column) of $\mathbf{K}$ be written as $K_i = [k(X_i, X_1), k(X_i, X_2), ..., k(X_i, X_N)]$.

2.5.2 Kernel as inner-product

For any kernel function $k(\cdot,\cdot)$ producing a positive semi-definite (PSD) kernel matrix $\mathbf{K}$, there exists a choice of basis functions $\phi(\cdot)$ such that $\langle \phi(X_i), \phi(X_j) \rangle = k(X_i, X_j)$. This is most simply due to the equivalence between PSD matrices and Gram matrices formed by inner products of vectors: a PSD matrix $\mathbf{K}$ has spectral decomposition $\mathbf{K} = V\Lambda V^\top$, and so $k_{i,j} = (\Lambda^{\frac{1}{2}} V_{[\cdot,i]})^\top (\Lambda^{\frac{1}{2}} V_{[\cdot,j]})$. Defining $\phi(X_i) = \Lambda^{\frac{1}{2}} V_{[\cdot,i]}$, we obtain $k_{i,j} = \phi(X_i)^\top \phi(X_j)$.[4]

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. The function space that is linear in these features can be understood through various intuitions, offered in Section 5.3.

2.6 Mean Balance on K

This section defines mean balance in terms of $\mathbf{K}$ and introduces useful notation. We 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 $\mathbf{K}$ among the treated can thus be denoted as $\overline{K_t} = \frac{1}{N_1} \mathbf{K_t}^\top \mathbf{1}_{N_1}$. Kernel balancing seeks weights that ensure the average row K_i of the treated is equal to the weighted mean K_i of the controls, which we term "mean balance on $\mathbf{K}$",

Definition 1 (Mean balance on $\mathbf{K}$) *The weights w_i achieve mean balance on $\mathbf{K}$ when*

$$\overline{K_t} = \sum_{i:D=0} w_i K_i$$

[4]The generalization of this to infinite-dimensional eigenfunctions is given by Mercer's Theorem (Mercer, 1909).

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

2.7 Replacing $\phi(X_i)$ with K_i

This section describes how the goal of achieving equal means on $\phi(X_i)$ for the treated and control can be replaced by the goal of achieving equal means only on the N-dimensional vectors K_i.

Consider fitting $\mathbb{E}[Y_{i0}|X_i]$ using models linear in $\phi(X_i)$, or equivalently, estimating θ in $Y_{0i} = \phi(X_i)^\top\theta + \epsilon_i$ with $\mathbb{E}[\epsilon_i|X_i] = 0$. One might fit such a model by regularized squared loss:

$$\underset{\theta\in\mathbb{R}^D}{min} \sum_i (Y_{0i} - \phi(X_i)^\top\theta)^2 + \lambda||\theta||^2$$

For any $\lambda > 0$, the resulting coefficients are representable as $\theta = \sum_i c_i\phi(X_i)$, which can be found either by directly seeking to minimize the regularized loss (see e.g. Hainmueller and Hazlett, 2014), or by appealing to the Representer Theorem (Kimeldorf and Wahba, 1970). Thus, accepting any non-zero degree of regularization, the model will always produce predictions of the form

$$\begin{aligned}\phi(X_i)^\top\theta &= \phi(X_i)^\top \sum_j c_j\phi(X_j) \\ &= \sum_j c_j\langle\phi(X_j), \phi(X_i)\rangle \\ &= \sum_j c_j k(X_j, X_i) = K_i c\end{aligned}$$

With $\mathbb{E}[Y_{0i}|\phi(X_i)] = K_i c$, we can instead use K_i as bases for the conditional expectation function rather than $\phi(X_i)$, which never even need to be constructed.

To review thus far, the intuition for the approach is that if we find weights w_i such that the (weighted) mean row K_i among the controls equals the (unweighted) mean among the treated, this also assures the weighted mean $\phi(X_i)$ among the controls equals the unweighted mean among the treated. Having assumed the linearity of Y_{0i} in $\phi(X_i)$, this suffices to ensure that Y_{0i} has the same mean in the two groups. This in turn ensures that differences in means or outcomes models run with those weights will be unbiased for the ATT.

2.8 Choice of weights: Approximate balance and resulting bias

What remains is to choose weights w_i to obtain balance on $\mathbf{K}$. As exact balance on all N dimensions of $\mathbf{K}$ is typically infeasible, we instead seek approximate balance. The approach taken here to approximate balance can be motivated by either considering a lower-dimensional approximation of $\mathbf{K}$ that can be balanced upon, or by seeking to minimize the (worst-case) bias due to imperfect balance. While the two are closely related, the former logic is described in Appendix A.4 while I focus here on the bias minimization approach.

In short, the procedure used here takes the eigenvectors of $\mathbf{K}$ by singular value decomposition, and achieves fine balance[5] on the first r of these, leaving eigenvectors whose eigenvalues rank $r+1$ to N unbalanced. The value of r is then chosen by a procedure described below that minimizes the worst-case bias in light of remaining imbalances and numerical limitations. In this section, I describe the motivation for this procedure, how it is implemented, the choice of dimension r, and the resulting worst-case bias due to this approximation.

2.8.1 Bias due to approximate balance

Here we directly examine the worst-case bound on the bias due to approximate rather than full balance on $\mathbf{K}$ as a means of motivating the eigen-decomposition approach to approximate balance.

Recalling Assumption 2, we have $\mathbb{E}[Y_{0i}|X_i] = \mathbf{K}c = \mathbf{V}\mathbf{A}\mathbf{V}^\top c = \mathbf{V}d$, where $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{K}$, $\mathbf{A}$ is the matrix whose diagonal contains the eigenvalues of $\mathbf{K}$, and d is a rewritten form of the "coefficients", c, that operate in the eigenvector space, with $d = \mathbf{A}\mathbf{V}^\top c$. Note also that the (Hilbert space) norm of this function is $c^\top \mathbf{K} c = c^\top \mathbf{V} d = d^\top \mathbf{A}^{-1} \mathbf{V} d$. Further, let $\mathbf{V}_1$ be rows of $\mathbf{V}$ corresponding to treated units, and $\mathbf{V}_0$ the rows of $\mathbf{V}$ corresponding to control units.

Suppose then we choose the vector of weights w_0 on the control units, and w_1 on treated units. Here, because we target the ATT, every element of w_1 is simply $1/N_1$. The bias of the ATT due to

[5]The term "fine balance" is used here when an attempt is made to achieve exact balance, but due to numerical limitations the result is in fact nearly-exact balance.

approximation, denoted $bias_w$, is then

$$bias_w = \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \tag{8}$$
$$= (w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)d \tag{9}$$
$$= (w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}\mathbf{V}^\top c \tag{10}$$

To obtain a worst-case bound on this bias when we do not know c (or d), we must instead control some related quantity. In particular, I propose imposing control over only the Hilbert norm of the regression function, $c^\top \mathbf{K} c$, as this controls how wildly the regression function is allowed to vary. Suppose we restrict the function to those with norm $c^\top \mathbf{K} c \leq \gamma$. We are then interested in the worst-case bias, due to approximation, *biasbound*, given by

$$\sup_{c^\top \mathbf{K} c \leq \gamma} |(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}\mathbf{V}^\top c|$$

Letting $z = c^\top \mathbf{K}^{1/2} \gamma^{-1}$, this can be rewritten as,

$$\sqrt{\gamma} \sup_{z^\top z \leq 1} |(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}\mathbf{V}^\top \mathbf{K}^{-1/2} z|$$

which by Cauchy-Schwarz gives

$$biasbound \leq \sqrt{\gamma} ||(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}\mathbf{V}^\top \mathbf{K}^{-1/2}||_2 \tag{11}$$
$$\leq \sqrt{\gamma} ||(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}^{1/2}||_2 \tag{12}$$

The form of this worst-case bound is informative. First, the L_2 norm of the regression function (γ) controls the overall scale of potential bias. Second, the imbalance on the eigenvectors of $\mathbf{K}$ after weighting, $(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)$, enters directly, and the contribution of each imbalance on the bias bound is that eigenvector's imbalance, times the square root of the corresponding eigenvalue. This scaling by eigenvalues suggests one approach to achieving approximate balancing weight: achieve fine balance on the first r eigenvectors, such that the imbalanced eigenvectors are only those with very small eigenvalues and thus of little consequence. Because the matrix $\mathbf{K}$ typically has a few large eigenvalues then many

very small ones (if the choice of b is appropriate), it is usually possible to achieve fine balance on enough eigenvectors such that the remaining eigenvalues carry a tiny fraction of the total variation in $\mathbf{K}$. This is the approach taken here, as described next.

Note that, while assuming a set of bases in which the non-treatment potential outcome is linear clearly introduces a cost, it provides several benefits. The first is that it tells us what functions of the covariates must be balanced across treatment status to imply unbiasedness, as is the premise of kernel balancing. Another more subtle upside is that, even when the balance achieved on these bases is approximate, having assumed linearity in these bases allows us to assume a worst-case bias due to the imperfections in balance.

2.8.2 Choice of r

What remains is the choice of r, and the procedure for finding the weights to balance r dimensions. We chose r so as to minimize this worst-case bias due to approximation, $||(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}^{1/2}||_2$, where $\sqrt{\gamma}$ has been dropped as it does not vary across choices of r. Weights are found to balance $r = 1$ dimensions, and the bias bound is computed. Then r is increased, until a choice of r that minimizes *biasbound* is found. This leaves the main imbalances on only the most inconsequential eigenvectors (i.e. those with small eigenvalues). While this method is thus motivated by minimizing the worst-case bias and appears to be effective in simulation and application, future work may fruitfully propose procedures to minimize $||(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}^{1/2}||_2$ in a different or more direct way through a choice of approximate balancing weights. In practice, the bias bound drops rapidly as r first grows and the major eigenvectors are swept up. This works up to a point, after which it becomes increasingly difficult to obtain balance on additional dimensions, and numerical instabilities in the optimization begin to accumulate, until finally no feasible solution exist. This is illustrated, together with other properties by simulation in Section 3.2.

2.8.3 Weight selection given r

So far we have described the constraints which a set of weights must satisfy – balance on the first r eigenvectors of $\mathbf{K}$ – but not how those weights are actually chosen. We have great flexibliity in the

choice of weights to achieve such constraints, and in particular, a measure of divergence from uniform weights we wish to keep at a minimum subject to achieving the balance constraints. Appendix A.1 describes implementation options consistent with the approach outlined here, and the particular choice implemented in the package `kbal`, which maximizes the entropy measure, $\sum_i w_i log(w_i)$, as suggested by Hainmueller (2012). A second choice (also implemented in the `kbal` software) is to use the weights that maximize the empirical likelihood subject to the balance constraints (Owen, 2001). This effectively maximizes $\sum_i log(w_i)$ subject to the constraints. Both methods work well here and the choice between them is beyond the scope of this paper. Alternative choices could also be made such as the minimum-variance weights described in Zubizarreta (2015) or the non-parametric covariate balancing weights described in Fong et al. (2018).

2.9 Alternative interpretation: smoothed multivariate balance

The principal motivation for kernel balancing is as a reliable and hands-off method for estimation of the ATT (or ATC or ATE, see Section 5.4) by obtaining equal means Y_{0i} for the treated and control groups as described above, under reasonable assumptions on $\mathbb{E}[Y_{0i}|X_i]$. However, the use of kernels for the choice of $\phi(X_i)$ above produces a very useful equivalence: kernel balancing using the kernel $k(\cdot,\cdot)$ implies that the multivariate density of the covariates *as estimated by the same smoothing kernel* $k(\cdot,\cdot)$ will be equal for the treated and control groups, at all covariate locations in the data. It thus also approximates in a finite sample the goal of "multivariate balance" normally targeted by matching and weighting procedures, but only insofar as those densities are well-estimated using that choice of kernel.

These multivariate density estimators may not be satisfactory density estimators as such, particularly in high-dimensional data. However, methods seeking multivariate density balance can typically only hope to achieve or verify that balance with respect to *some* density estimator or sample statistics, making this a very useful equivalence. As a corollary, a researcher seeking multivariate density balance could first commit to a kernel smoother she would would be willing to use to estimate the multivariate density in each group, after which kernel balancing produces the weights resulting in equality of these estimated densities.

Proposition 1 (Balance in **K** implies equality of smoothed multivariate densities) *Consider a den-*

sity estimator for the treated, $\hat{p}_{X|D=1}$ and for the (weighted) controls, $\hat{p}_{X|D=0,w}$, each constructed with kernel $k(\cdot,\cdot)$ of bandwidth b as described below. The choice of weights that ensures mean balance in the kernel matrix $\mathbf{K}$ ensures that $\hat{p}_{X|D=1} = \hat{p}_{X|D=0,w}$ at every position at which an observation is located.

Proof of Proposition 1 is given in Appendix A.6. Here I briefly build an intuition for this result, as it leads to further insights and tools. First, the typical Parzen-Rosenblatt window approach estimates a density function according to:

$$\hat{p}(x) = \frac{1}{N\sqrt{4\pi b}} \sum_{i=1}^{N} k(x, X_i) \tag{13}$$

for kernel function $k(\cdot,\cdot)$ with bandwidth b. The Gaussian kernel is among the most commonly used for this task. While typically considered in a univariate context, Expression 13 utilizing a Gaussian kernel generalizes to a multivariate density estimator based on Euclidean distances.

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, $\frac{1}{N\sqrt{4\pi b}}$, 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 $\frac{1}{N_1\sqrt{4\pi b}}\mathbf{K_t}^\top \mathbf{1}_{N_1}$ where $\mathbf{1}_{N_1}$ is a N_1-vector of ones thus returns a vector of estimates for the density of the treated, as measured at all observations. Finally, $\frac{1}{N_0\sqrt{4\pi b}}\mathbf{K_c}^\top \mathbf{1}_{N_0}$ returns estimates for the density of the control units at every datapoint in the sample, and $\frac{1}{\sqrt{4\pi b}}\mathbf{K_c}^\top w$ gives the w-weighted density of the controls, again as measured at every observation.

Using the formulas above, we can rewrite the estimated density of the treated as $\frac{1}{N_1\sqrt{4\pi b}}\sum_i K_i$ and the weighted estimated density of the controls as $\frac{1}{\sqrt{4\pi b}}\sum_i K_i w_i$. Setting these equal to each other gives

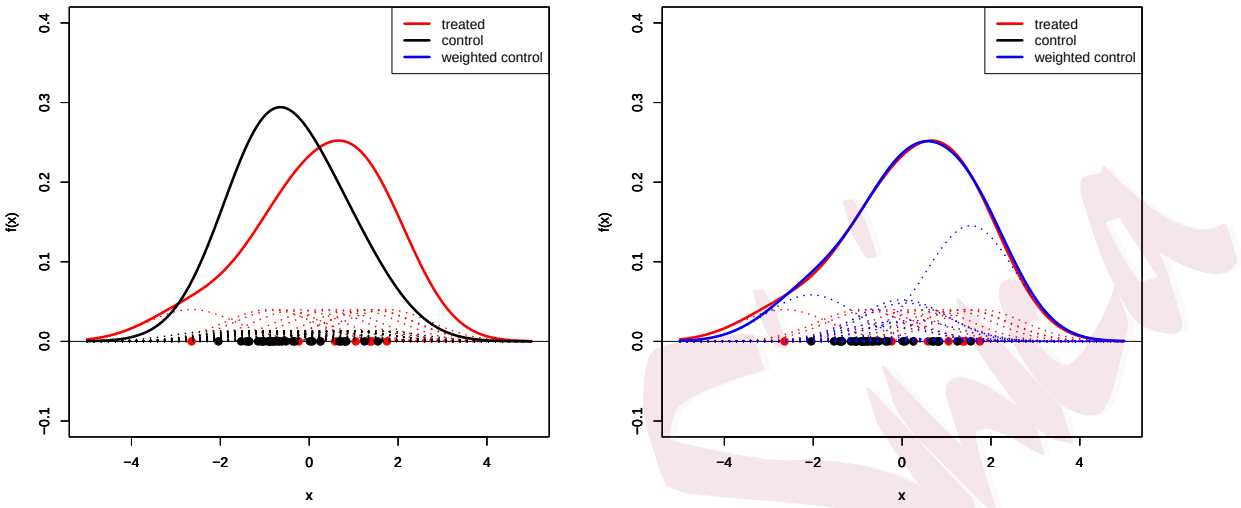
$\frac{1}{N_1}\sum_i K_i = \sum_i K_i w_i$ – which is simply the mean balance on K_i condition approximately satisfied by kernel balancing. This illuminates the deep connection between (i) an assumption one makes on the outcome space of models and (ii) the smoother for which estimated multivariate density balance is achieved.

Of greatest practical relevance to investigators, this suggests a measure of imbalance relating to the difference between the kernel-estimated distribution of covariates for the treated and for the control, both before and after weighting. Let us consider first our goal of achieving mean balance on K_i. To minimize smoothed multivariate imbalance under this kernel, we might wish to minimize a p-norm proportional to $||\overline{K_t} - \sum_{i:D=0} w_i K_i||_p$. On the other hand, we may wish to think of the implied estimate for the density of the treated at every observation's covariate location, and the implied estimate for the density of the controls at each location, and take the "difference in heights" between them, as in $\frac{1}{2}||\hat{p}_{D=1}(\mathbf{X}) - \hat{p}_{w,D=0}(\mathbf{X})||_p$. Fortunately, we need not choose, as the latter is equal to $\frac{1}{2}||\frac{1}{N_1\sqrt{4\pi b}}\mathbf{K_t}^\top \mathbf{1}_{N_1} - \frac{1}{\sqrt{4\pi b}}\mathbf{K_c}^\top w||_p$, and is thus the same as the first. See Appendix A.1.2 for details.

The L_1 norm, $\frac{1}{2}||\frac{1}{N_1\sqrt{4\pi b}}\mathbf{K_t}^\top \mathbf{1}_{N_1} - \frac{1}{\sqrt{4\pi b}}\mathbf{K_c}^\top w||_1$, is useful as it is naturally interpertable as an average of the gap between the kernel-estimated density of the treated and control at every observation, each scaled to properly integrate to 1. This is analogous to the L_1 norm proposed by (Iacus et al., 2011) for use with Coarsened Exact Matching, but here does not require coarsening the covariates into discrete bins as proposed there. However, because this interpretation holds only insofar as the implied kernel density estimator is a good estimator, it should be used with caution. Note that the L_1 norm is closely related to *biasbound*, and has extremely similar behavior as a function of r (see Section 3.2).

Figure 1 illustrates 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}, through mean balance on a suitable set of bases (K_i), that is essential for proving unbiasedness, gives rise to the bias bound and other analytical results, and which kernel balancing targets.

Figure 1: 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 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

3 Simulation Examples and Evidence

3.1 An Illustration: Imbalance on a ratio

Building on the simple example given in Section 1, this simulation highlights the practical challenges of existing methods and demonstrates the effectiveness of kernel balancing against these challenges. To keep real world relevance in mind, rather than using generic variables names such as X_1 or Y, I describe it in terms of a realistic example, where the proposed multivariate confounder could easily be imagined to exist.

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, within the set of civil war cases constituting our sample, 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, suppose the investigators collects 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 estimating an ATT, defined as the expected number of $peace\,years$ experienced by countries that received $peacekeeping$, minus the expected number of $peace\,years$ for this group had they not received peacekeeping missions.

Further, suppose there are no unobserved confounders, and that peacekeeping missions are deployed only on the basis of these observables – but not necessarily linear functions of them. Specifically, suppose that a conflict's $intensity$, which equals $\frac{fatalities}{war\,duration}$ is the key confounder that relates to both treatment assignment ($peacekeeping$) and the outcome ($peace\,years$). In particular, missions are more likely to be deployed where conflicts were higher in intensity, with treatment assigned by:

$$peacekeeping_i \sim Bern(logit^{-1}\left(\frac{intensity}{5000} - 1\right))$$

with $war\,duration$ distributed as $max(1, N(7, 9))$ and $intensity$ in fatalities per year distributed Unif$(10^2, 10^4)$. The observed covariate $fatalities$ is constructed according to $intensity \cdot war\,duration$.

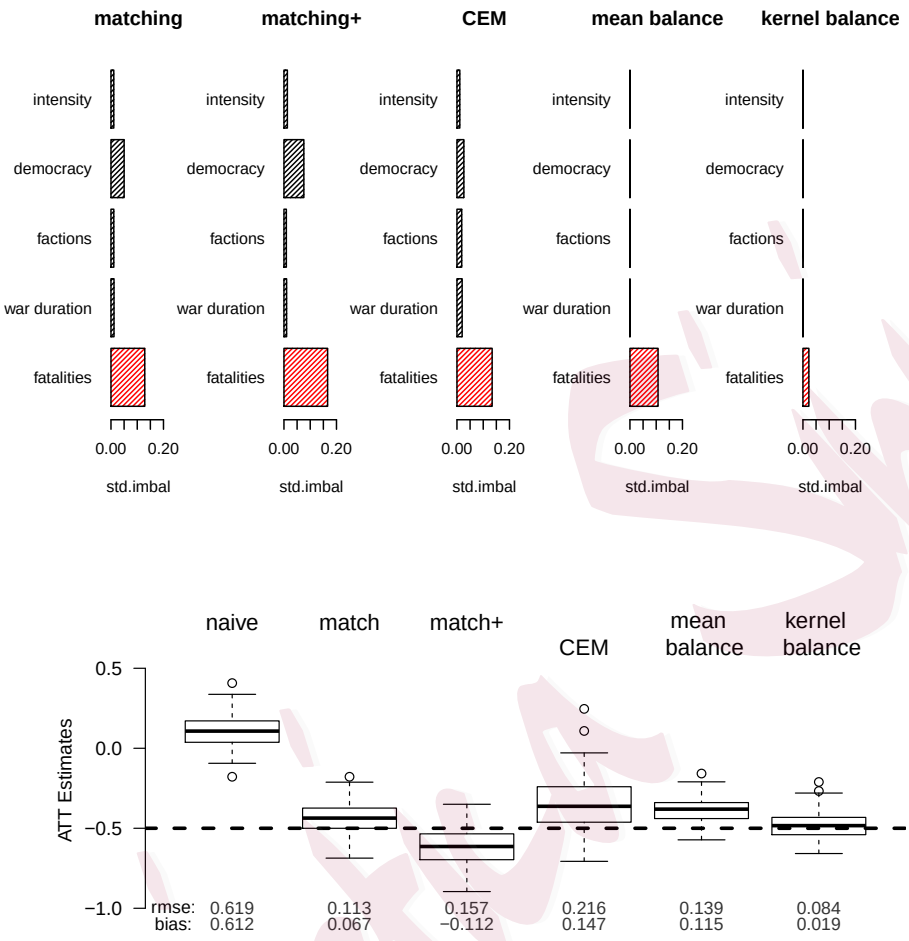
Suppose the outcome of interest, $peace\,years$, is also a function of $intensity$, with more intense conflicts leading to longer average $peace\,years$ according to

$$peace\,years = 5 + 2\frac{intensity}{5000} - (0.5)peacekeeping_i + \epsilon_i$$

which generates a fixed treatment effect of -0.5 years, and ϵ_i is an error term drawn from $N(0, 4)$. Such a scenario, in which $intensity$ is positively associated with both the probability of receiving a peacekeeping mission (treatment) and more years of peace (the outcome) is plausible. For example, more intense wars are more likely to attract the attention of the international community and result in deployment of a mission, but may also indicate greater dominance by one party to the conflict, leading to a lower likelihood of resurgence in each subsequent year.

How well do existing techniques achieve equal means for the treated and controls ("mean balance"), both on the original four covariates and on $intensity$, a (non-linear) function of the observables? In

Figure 2: Simulation: Imbalance on a Ratio



Results from 500 simulations of the peacekeeping example described in the text. The methods employed are: (*Matching*), one-to-one Mahalonobis distance matching with bias adjustment; (*matching+*), matching on full second order expansion of the covariate (14 terms in total) with bias adjustment; (*CEM*), coarsened exact matching at default values; (*mean balance*), entropy balancing weights for equal means on the observed covariates; and (*kbal*), kernel balancing at the default settings. *Top*: Standardized covariate imbalance by method. All methods except kernel balancing (*kbal*), achieve poor balance on the unknown but important function of observables, *intensity*. *Bottom*: boxplots illustrating distribution of average treatment effect on the treated (ATT) estimates. The actual effect is -0.5 *peace years*. All methods except for kernel balance show large biases in the ATT estimates, which arise due to the persistent imbalance on *intensity*.

Figure 2, the top panel shows covariate imbalance on the horizontal axis (the standardized difference in means between treated and control), on each of the covariates as well as the key function of covariates, *intensity*. All results are taken over 500 simulations with the same data generating process and $N = 500$ each time. 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 2 (*matching+*) shows that even this approach would not generate balance on *intensity*. In fact, balance on both *war duration* and *intensity* are worsened. Next, Coarsened Exact Matching (CEM) is used, which coarsens variables so that exact matching on the resulting data is possible (Iacus et al., 2011). This also does not solve the problem: imbalances remain on the original, uncoarsened variables. Fourth, (*mean balance*) employs entropy balancing (Hainmueller, 2012) 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, the fifth plot shows results from *kernel balance*. Because this method achieves balance on many smooth functions of the included covariates, it achieves vastly improved balance on *intensity*.

These imbalances are worrying precisely because they indicate a failure to condition on the covariates as intended. Because the imbalanced covariate directly influences the potential outcomes here, this imbalance leads to biased ATT estimates. The resulting ATT estimate for each method is shown in the bottom panel of Figure 2. Large biases occur for each estimator, with the exception of kernel balancing. Note that for both *matching* and *mathing+*, bias-adjustment as per Abadie and Imbens (2011) is employed in an effort to make up for matching discrepancies on the observed covariates. This however cannot account for the non-linear effects of observables on the potential outcomes. Kernel balancing shows the lowest bias by far among the methods attempted. It's advantages in RMSE are more modest, but it still has 22% lower RMSE than the next best estimator, *mean balance*.

Though kernel balancing is largely automated, given a choice of Gaussian kernel one still chooses the bandwidth parameter, b. Section A.10 descibes the substantive meaning of this parameter, but it is useful to examine the sensitivity of results to choices of b. Figure 8 in the Appendix shows that estimates are stable across choices of b ranging from one quarter to four times the default choice of $dim(X) = 4$ (see Section A.10 for discussion of this default value).

This illustration demonstrates the ease with which existing methods may fail: when a confounder is

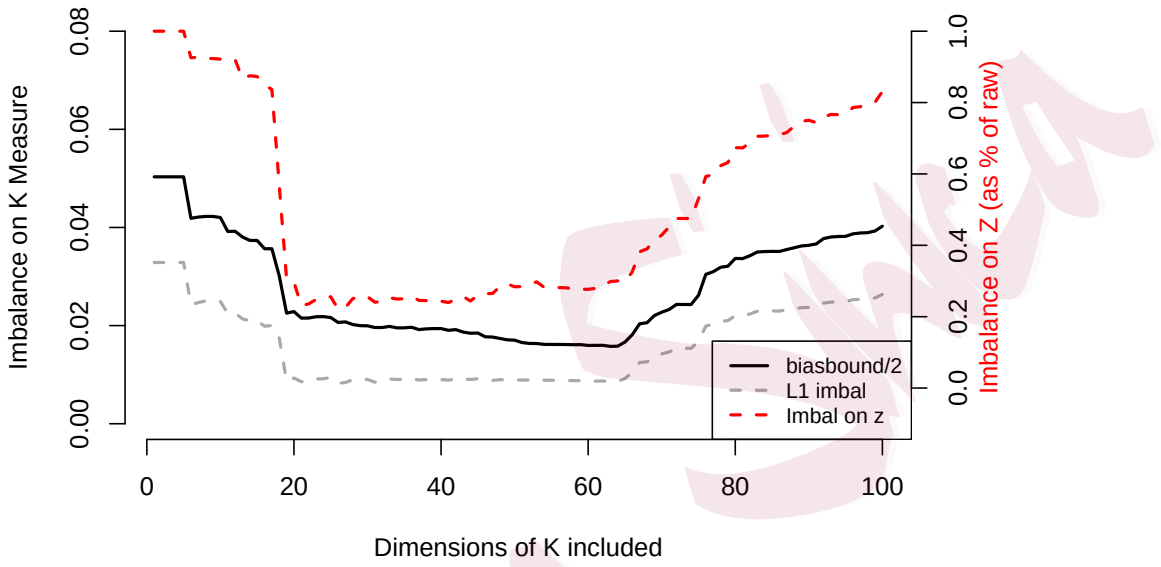
a non-linear function of two observed covariates, even a simple ratio, existing matching and weighting methods pose risks of large biases. An investigator's theoretical knowledge is rarely sufficient to ensure the investigator can guess what functions of the observables may impact the outcome. Kernel balancing provides one principled approach for choosing function of the covariates on which to achieve balance to ensure unbiased estimation in a wide range of plausible scenarios – those where the non-treatment potential outcome is a smooth function of X_i. An illustration of the effectiveness of the worst-case bound in simulations is given in Appendix A.3.

3.2 Behavior of Bias Bound and L_1 imbalance across r

In this simulation, we examine how the bias bound and the L_1 imbalance vary together as we increase the number of eigenvectors, r that enter the balancing procedure. We also examine whether the bias bound (or L_1) provides an accurate guide to minimizing imbalance on unknown functions of the covariates by explictly forming such a function and checking imbalance on it at each step.

Let $x_{1i}, ..., x_{5i}$ be covariate data, each drawn $N(0, 1)$ for $i \in 1, ..., 500$. Let $z_i = \sqrt{x_{1i}^2 + x_{2i}^2}$. This function impacts treatment assignment, with probability of treatment being given by $logit^{-1}(z_i - 2)$, producing approximately two control units for each treated unit. In Figure 3, the value of r – the number of factors of $\mathbf{K}$ retained for purposes of balancing – is increased up to 100. The bias bound shown here is not scaled – i.e. it is computed as if $\gamma = 1$ to illustrate how it changes across r given a constant choice of γ (it is then divided by 2 for graphical convenience). As expected, the bias bound, the L_1 score, and the mean imbalance on z after weighting improve rapidly as r is first increased, with the most important eigenvectors coming into balance. It then plateaus, and eventually worsens beyond some choice of r. Most importantly, while the balance on z is unknown to the investigator, the bias bound and L_1 are observable, and improvements in balance on z track well with them. Accordingly, selecting r to minimize the bias bound appears to be a viable strategy for selecting the value that also minimizes imbalance on unseen functions of the data in this example. As also expected, the bias bound and L_1 are very similar, up to a scaling factor – in fact, all three quantities in Figure 3 correlate with each other above 0.98. Note also that there is a wide range of r values – approximately 20 to 60 – that produce similar levels of imbalance, making the exact choice less critical.

Figure 3: Choice of r: bias bound, L_1 imbalance, and imbalance on an unknown function of the observables



Imbalance measures over values of r, the number of dimensions balanced. The L_1-imbalance score is interpretable as the L_1 measure of the gap between the estimated densities of the treated and control covariates, when that approximation is made by the same kernel function used to form $\mathbf{K}$. *biasbound* is the derived worst-case bound on the bias due to the approximate nature of balance (the *biasbound* is divided by 2 simply for graphical convenience). The actual bias bound would further be scaled by the choice of Hilbert norm for the outcome model, but here this is irrelevant as we examine variation across the choice of r. Finally, $z = \sqrt{x_1^2 + x_2^2}$ is a function of the observable covariates, which unknown to the investigator, may be confounding. Both the L_1 and *biasbound* closely follow the imbalance on z, such that choosing r to minimize either L_1 or *biasbound* is a sensible strategy for achieving minimum imbalance on z.

4 Example: National Supported Work Demonstration

It is useful to know whether kernel balancing accurately recovers average treatment effects in observational data under conditions in which an experimental benchmark is available for comparison. This can be approximated using a method and dataset owed to LaLonde (1986) and Dehejia and Wahba (1999), and which has become a routine benchmark for matching and weighting approaches in a disciplines as diverse as statistics, econometrics, political science, psychology, and epidemiology (see e.g. Diamond and Sekhon, 2005; Iacus et al., 2011; Hainmueller, 2012; McCaffrey et al., 2004; Little and Rubin, 2000). 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. For this treated group, the experimental estimate of the ATT is \$1794. For the observational version of the study, we keep these treated units, but draw the control sample from the Panel Study of Income Dynamics (PSID-1) containing 2490 individuals.

The pre-treatment covariates available are age, years of education, an indicator for no highschool degree, real earnings in 1974, real earnings in 1975, indicators for zero income (taken to mean unemployment) in 1974 and 1975, and a series of demographic indicators variables: black, hispanic, and married. 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, not to mention the constructed indicators for zero income in 1974 and 1975. 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. The general supposition of kernel balancing, however, is that investigators would not typically know (or be expected to know) that these particular non-linear transformations are required.

In this reanalysis, three estimation approaches are compared, with three specifications attempted for each. The first procedure used is simply linear regression (OLS), which is not an effective competitor but rather serves to show that assuming a simple outcome model can produce highly problematic results. Second, Mahalanobis distance matching ($match$) is employed, with bias adjustment. Third,

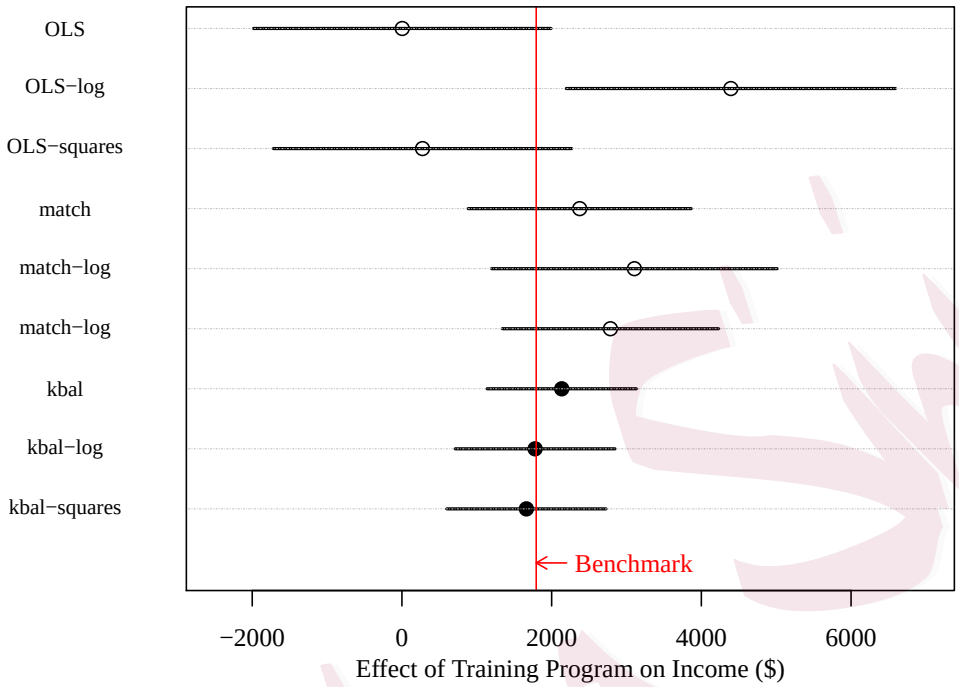
kernel balancing (*kbal*) is used, with b set to the default value of P (the number of covariates). For comparability, all three approaches use simple standard errors that take the weights as fixed.

For each method, three specifications are attempted, chosen on the grounds that they are reasonable choices we might expect investigators to make and justify in their analyses. First, the "standard" set of ten covariates described above. Second, a reasonable investigator might propose that log income is a better choice than raw income for determining who should be considered similar and be matched together, and so incomes in 1974 and 1975 are replaced with their logs (plus one). Third, a thoughtful investigator may instead be concerned about flexible functional forms for the reasons raised in this paper, and try expanded set of covariates including the ten standard covariates, plus the squared terms for the three of these that are continuous. Note that all three approaches seem reasonable enough as to not necessarily raise suspicion.

Figure 4 shows that the OLS estimates are poor and vary widely by specification. This reflects that large differences between the locations of treated and control units in the dataset, and that using a linear model to account for these imbalances fails. Mahalahobis distance matching performs much better, though remains somewhat specification dependent, with its best estimate (*match-squares*) falling within $581 of the benchmark. Finally, kernel balancing (*kbal*) performs well across all three specifications, with no estimate more than $341 from the benchmark. Whether one constructs additional squared terms, or believes they should take the log of income, the space of functions represented in the span of $\mathbf{K}$ is large and flexible, so the resulting solutions change little. The kernel balancing estimates are closer to the benchmark in each case, less sensitive to specification, and simultaneously show less uncertainty. This reduced variance relative to other methods is likely due to the improved finite sample balance on characteristics influencing the outcome (see Ho et al., 2007 for related discussion on improved efficiency due to pre-processing). The resulting ATT estimates are very stable to the choice of b, once b exceeds a minimal value (See Figure 9 in Appendix).

The resulting weights can yield insights. At the solution achieved by kernel balancing using the default value of $b = 10$, 90% of the weight for controls is taken from just 85 units. Relatedly, we see the that initial sample was badly imbalanced, with an L_1 distance of 45%. Fortunately, weights can be found to eliminate this, as the L_1 distance drops to 0.1% after weighting. Looking to the 85 highly

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*) with bias adjustment, and kernel balancing (*kbal*). For each, three sets of covariates are attempted: the standard set of 10 covariates described in the text, a version replacing income in 1974 and 1975 with log income (*log*), and an expanded set (*squares*) including the 10 standard covariates plus squares of the three continuous variables. The experimental benchmark of $1794 is indicated by the vertical line. While both *match* and *kbal* produce reasonable results, *kbal* results are closest benchmark, showing the least sensitivity to specification.

weighted units, we see they are predominantly the unemployed. This points us to the main source of imbalance: while 72% of the treated are unemployed in either 1974 or 1975, only 12% of controls are unemployed in either year. In fact, one need *only* seek balance on unemployment in 1974 and in 1975 in order to find reasonable results.[6]

[6]An additional worked application is included in Appendix (A.12), which applies kernel balancing to re-examine the results of (Lyall, 2010) regarding whether democracies are less successful in fighting counterinsurgencies.

5 Discussion

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

5.1 Relation to Existing Approaches

The most widespread tools to which kernel balancing can be compared include matching, covariate balancing or calibration weights, and propensity score methods. I also briefly contrast the approach to the more traditional strategy of simply fitting an outcome model in a suitable space of functions.

5.1.1 Matching

Under conditional ignorability as defined in Assumption 1, sub-classification and exact matching estimators conditioning on X and average the resulting estimates 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.

However, conditioning on X in this way is impossible when X is continuous or contains indicators for many categories, since we cannot literally compute differences 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 can remain incomplete, and the distribution of X for the treated and controls will not be made identical. The resulting bias in (S)ATT estimates dissipates only very slowly as N increases, and in general

the resulting estimates are not $\sqrt{N}$-consistent (Abadie and Imbens, 2006). To minimize bias due to remaining matching discrepancies, investigators are sometimes instructed to attempt different matching specifications and procedures until they achieve satisfactory multivariate balance (see e.g. Stuart, 2010). However in practice, tests for this balance are usually limited to univariate tests comparing the marginal distribution of each covariate under treatment and control. As the simulation example in Section 3.1 illustrates, many matching approaches can thus fail to obtain sufficient similarity of distributions, even when investigators attempt to match on higher-order terms. Relatedly, a class of methods referred to as *optimal* matching minimizes a global measure of distance between the distributions of treated and control units. Kallus (2016) consider a generalization of optimal matching methods as those that chose a bias-variance tradeoff so as to minimize the worst-case conditional mean squared error, and show that many existing techniques are optimal in this sense depending on assumptions on the conditional expectation of the outcome and the residual variance. Further, they propose Kernel Optimal Matching, which solves a minimization problems involving a Guassian or other kernel representation of the data in the minimization objective. This proves to have many useful properties, especially when paired with an outcome regression model, suggesting another route by which kernels may be useful in procedures to estimate treatment effects.

5.1.2 Covariate Balancing Weights

Another category of methods for multivariate balancing is 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 (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; Fong et al., 2018), with a number of related procedures emerging earlier in 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 enough to complete the adjustment. 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 often be made exactly equal to those of the treated, avoiding the matching discrepancy

problem. The downside is that it sacrifices 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 of these covariate balancing weighting methods that choices what moments to ensure balance on by constructing bases whose span provides a flexible space for the outcome model.

5.1.3 Propensity Score Weighting

Propensity score methods such as inverse propensity score weighting can similarly be understood as an attempt to find the weights that make the distribution of the covariates for the controls and treated similar (in expectation), but through adjusting for estimated treatment probabilities.

For purposes of ATT estimation, the stabilized inverse propensity score weights applied only to the control units would be $w_{IPW} = \frac{p(D_i)}{p(D_i|X_i)} \frac{1-p(D_i|X_i)}{1-p(D_i)}$. Appendix A.8 shows how these weights can be derived as those that transport the distribution of the controls to match that of the treated during ATT estimation. Moreover, these weights can be rewritten via Bayes rule as the ratio of class densities for the treated and controls,

$$w_{IPW} = \frac{p(x|D_i = 1)}{p(x|D_i = 0)} \tag{14}$$

Written in this way, it becomes clear that whenever the class densities are equal for the two groups, the IPW weights on controls for ATT estimation would be constant. Given the multivariate balancing property discussed above, kernel balancing weights approximately achieve this equality but with the *estimated* class densities corresponding kernel density estimator (Section 2.9). Alternatively, suppose one estimated the propensity score with a generative classification model, in which the class densities for the treated and controls are estimated using kernel k as a smoother. If the resulting inverse propensity score weights are constructed so as to estimate the ATT, the result will equal that from kernel balancing, up to the approximation based on r.

5.1.4 Comparison to Outcome Models

An alternative and common estimation route is simply to regress the observed Y_i on some (possibly augmented) set of covariates X_i and treatment D_i. How to combine the power and flexibility of machine learning or high dimensional models, with an outcome model that efficiently and unbiasedly returns estimates of causal effects, remains a very active area of research. Regularized regression models have been employed to accomodate high dimensional covariates, however the shrinkage imposed by these models leads to substantial biases and poor inferential properties (Belloni et al., 2014). A series of doubly-robust or debiasing methods utilizing a (consistent) estimator of the propensity-score to adjust these models have been proposed, following Robins et al. (1994) (see e.g. Van der Laan and Rose, 2011; Farrell, 2015; Belloni et al., 2017; Chernozhukov et al., 2017). Further recent efforts have sought to avoid the requirement of a consistently estimated propensity score model to make such adjustments. For example, Ratkovic and Tingley (2017) propose a Bayesian sparse model for variable selection that, combined with feature expansions such as a tensor-spline, performs well and can be easily extended to other approaches. Athey et al. (2016) effectively combine the outcome model approach with weights that seek covariate balance, by reweighting residuals from a linear model using weights that achieve (mean) balance on covariates. Like these methods, kernel balancing adopts insights from machine learning, relying on properties of kernels. However, unlike other work importing machine learning methods, it does not use them to solve a classification or regression problem such as fitting the outcome or a propensity score. Rather, it uses kernels to establish a high-dimensional choice of bases, which tell us what functions of the covariates must be balanced.

Two important distinctions can be made between assuming an outcome model for purposes of choosing "what to balance on" versus fitting an outcome model. The first is that kernel balancing works regardless of *which* function in the function space is the correct one (i.e., the value of θ or $\mathbf{c}$) – we do not need to rely on the accuracy of estimates for these coefficients in a finite sample. We need only that such a model exists, and even then, violations of the model are bias-inducing only in certain cases (see Appendix A.2.1). Second, employing a weighting approach whose justification is rooted in a choice of outcome models is not equivalent to using the outcome model alone, because (when estimating the ATT) the former changes the distribution of the control group to be more similar to that of the

treated prior to estimation of an effect. Such a "pre-processing" approach reduces model-dependency, avoiding heroic modeling assumptions to bridge the gap between treated and control units that may lie far apart in the covariate space (see e.g. Ho et al., 2007 for analogous arguments in the matching literature).

5.2 Uncertainty Estimation

In most contexts, investigators require a measure of uncertainty such as a standard error or confidence interval around their effect estimates. With matching estimators, one approach is to ignore the uncertainty due to the matching procedure itself. For example Ho et al. (2007) argue that since variance estimators for parametric models typically take the data as fixed anyway, when data are pre-processed by a matching procedure, the matched dataset can be taken as fixed for subsequent analyses as well. Thus, the variance can be estimated for parametric outcome models on the matched data in the usual way, i.e. by applying weights that reflect which control units are dropped or multiply used to the outcome model of interest and computing the associated standard errors. Similarly, weighting estimators such as entropy balancing may also take this pre-processing view and treat the resulting weights as fixed (Hainmueller, 2012) for purposes of computing uncertainty estimates in subsequent analyses.

In contrast, Abadie and Imbens (2008) consider the uncertainty due to the matching process, noting that the bootstrap fails in this case due to the "extreme non-smoothness" of matching estimators. Abadie and Imbens (2006) develop asymptotic standard errors that account for uncertainty in the matching procedure. Others have argued that an m-out-of-n bootstrap may be appropriate (see Politis and Romano, 1994). One benefit of kernel balancing and other weighting methods is that, because the weights are continuous and observations are not wholly dropped as in matching, the simple bootstrap may be valid. However, the development of more computationally attractive alternatives remains an area for ongoing research.

5.3 Gaussian kernel and intuition for $\phi(X_i)$

One reason to use the Gaussian kernel is that it is the workhorse kernel in machine learning regression and classification tasks. Though kernel balancing does not actually fit an outcome model here, the

function space invoked, $\phi(X_i)^\top\theta$, is the same as that in which kernel methods such as kernelized regression, support vector machines with kernels, and Gaussian processes operate. The Gaussian kernel has the universal representation property: as $N \to \infty$, $\phi(X)^\top\theta$ it encompasses any continuous function of X (Micchelli et al., 2006). While asymptotically appealing, this universality as N approaches ∞ is less reassuring in finite samples. However, smoother functions can be well 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. In many settings, such smoothness is reasonable: we expect that small changes in X_i should lead to small changes in Y_{0i} for the most part.

One approach to better understanding this function space is to analyze the features, $\phi(\cdot)$, consistent with the Gaussian kernel, i.e. so that $\langle\phi(X_i), \phi(X_j)\rangle = k(X_i, X_j)$. Since the choice of $\phi(X_i)$ implied by the Gaussian kernel is infinite-dimensional, it may seem difficult to imagine what this function space looks like. One valid choice for $\phi(X)$ in the case of the Gaussian kernel (with one dimensional X) is the sequence given by $\left\{\sqrt{\frac{2^d}{d!}}exp(-X_i^2)(X_i)^d\right\}$ for $d = 0, 1, ..., \infty$. Appendix A.7 describes this in greater detail, but fortunately a more intuitively useful understanding of this function space is available. As shown above the functions linear in $\phi(X_i)$ are also those linear in K_i. Accordingly, $k(X_i, \cdot)$ is sometimes called the "canonical feature mapping" corresponding to $\phi(x)$, (e.g. Minh et al., 2006). Because $k(X_i, x)$ evaluates at x the height of a Gaussian that had been centered at X_i, this function space is that which can be built by superposition of Gaussians placed over each observation and arbitrarily rescaled. That is, in the original covariates space $\mathbb{R}^P$, suppose we place a p-dimensional Gaussian kernel over each observation in the dataset, rescale each of these by a scalar c_i, then sum these rescaled Gaussians to form a single surface. By varying the values of c_i, 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 smooth, highly non-linear functions.

Another key question in determining what kernel to use is the choice of bandwidth, b. A useful default value for b is given by the column rank of X (see Appendix A.10 for explanation). An easy and transparent guideline for investigators is to show results using a range of choices for b, starting

from one half or less of the default value ($\dim(X_i)$) up to several times that value. Further details regarding the choice of b and its implications are discussed in Appendix A.10. The stability of results over choices of b in both the simulation and applied example here is illustrated in Appendix A.11.

5.4 Other Quantities: ATE, ATC

This paper has focused on the ATT for simplicity of exposition and comparability with matching approaches which often focus on the ATT as well. With minor adjustment, this method can also be used to identify the average treatment effect on the controls (ATC), and the average treatment effect (ATE).

To estimate the ATC, informally speaking we wish to "move the treated to the control locations" instead of the other way around. Accordingly, we instead seek weights on the treated units such that the weighted sum of K_i among the treated equals the (unweighted) average among the controls. That is rather than seeking the non-negative weights summing to one such that $\overline{K_t} = \sum_{i:D=0} w_i K_i$, we would instead seek the weights:

$$\overline{K_c} = \sum_{i:D=1} w_i K_i, \; \sum_i w_i = 1 \text{ and } w_i > 0$$

where $\overline{K_c}$ is the empirical average K_i taken over the controls only.

Similarly, for the ATE the goal is to transport both the treated and control to the same location and (more importantly) the same expectation of Y_{0i}. Thus we would seek the weights $w_i^{(1)}$ on the treated and $w_i^{(0)}$ on the controls such that

$$\sum_{i:D=0} w_i^{(0)} K_i = \sum_{i:D=1} w_i^{(1)} K_i = \overline{K}$$

where $\overline{K}$ is the empirical average of K_i taken over all the observations, treated and control alike. The `KBAL` package estimates the ATT by default but optionally estimates the ATC and ATE as well.

6 Conclusions

In the ongoing quest to reliably infer causal quantities from observational data, the primary challenge is often in ensuring that a set of observables sufficient for achieving ignorability can be found and that unobserved confounding can be ruled out. However, even then, the mechanics of actually performing the required conditioning on observables – particularly with multiple, continuous covariates – 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 systematically influences the non-treatment outcome persists in having a different mean for the treated and controls, the resulting estimates may 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. As illustrated here, when even a simple ratio of observables is confounding, existing methods can widely fail to complete the desired adjustment.

Fortunately, unbiasedly estimating the ATT requires only that the expected non-treatment potential outcome is equal in the treated and control group after adjustment. Kernel balancing seeks to achieve this by first assuming $\mathbb{E}[Y_{0i}|X]$ falls in a large space of smooth functions, which is in turn linear in the columns of 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 approximately equal to the average row of $\mathbf{K}$ for the treated. This ensures that the expected non-treatment outcome is approximately equal in the two groups. The remaining scope for bias due to the approximate nature of the weights can be bounded, and the weights are chosen by a method that seeks to minimize this worst-case bias. An alternative interpretation of the procedure is that kernel balancing implies that a particular kernel-based smoother for the multivariate densities is approximately equal for the treated and control, as evaluated at every observation.

Numerous extensions remain for future work. First, $\mathbf{K}$ has dimensionality $N \times N$, which becomes unwieldy as N grows large, posing a practical limit of tens of thousands of observations. Second, while the bootstrap may provide confidence intervals that include uncertainty due to weight selection, further work on this is needed, particularly on approximations that may not be as computationally burdensome when N is large. Finally, improvements may be possible on a number of implementation

details, such as the choice of b, the optimization procedure for choosing the number of dimensions and alternate methods for achieving approximate balance, and on understanding the potential for bias due to imperfect balance through metrics less extreme than the worst-case bound. An implementation of this procedure using the choices described here is available in the `R` package `KBAL`.

References

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

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1).

Athey, S., Imbens, G. W., and Wager, S. (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464.

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.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

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*, pages 779–801.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Fong, C., Hazlett, C., Imai, K., et al. (2018). Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177.

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.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

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.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

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

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

King, G., Lucas, C., and Nielsen, R. A. (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2):473–489.

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*, pages 604–620.

Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, 21(1):121–145.

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

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446.

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

Minh, H. Q., Niyogi, P., and Yao, Y. (2006). Mercers theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer.

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

Owen, A. B. (2001). *Empirical likelihood*. Wiley Online Library.

Pearl, J. (2009). *Causality*. Cambridge university press.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.

Ratkovic, M. and Tingley, D. (2017). Causal inference through the method of direct estimation. *arXiv preprint arXiv:1703.05849*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 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*, pages 472–480.

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

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.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1):305–353.

Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1923 [1990]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

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.

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

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

A Online Appendix

A.1 Further Implementation Details

A.1.1 Choice of Discrepancy Measure

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, by some measure, and in particular to avoid large weights. Two natural candidates for this are empirical likelihood (Owen, 1988), and entropy balancing (Hainmueller, 2012), both special cases of Cressie-Read divergence from a uniform distribution (Cressie and Read, 1984). 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. In the `kbal`, I utilize entropy balancing, which seeks to satisfy these conditions while maximizing the Shannon entropy, $\sum_i w_i log(w_i)$, implied by the weights, which is also (proportional to) the Kullback divergence entropy between the distribution of weights and a uniform distribution. See Hainmueller (2012) and references therein for further discussion.

A.1.2 Equivalence of K-imbalance and smoothed multivariate density imbalance

Recall that the choice 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 $\frac{1}{\sqrt{2\pi b}}$ to obtain $||\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t}^\top \mathbf{1}_{N_1} - \frac{1}{\sqrt{2\pi b}}\mathbf{K_c}^\top w||$ we see this equals $||\hat{p}_{D=1}(\mathbf{X}) - \hat{p}_{w,D=0}(\mathbf{X})||$, a norm on the difference between the smoothed density estimators for the treated and (weighted) controls, evaluated at each observation in the dataset. Hence, norms of the form $||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ are especially useful to minimize during optimization, as done in the selection of r here, because they both minimize imbalance in $\mathbf{K}$ and a reasonable measure of "multivariate imbalance", i.e. a norm over the different in multivariate densities for the treated and control.

When interpreted as a difference between estimated densities, the L_1 version of this norm described above is very much analogous to the L_1 metric used in Coarsened Exact Matching (Iacus et al., 2011), but without requiring coarsening in order to construct discrete bins in the covariates space.

A.2 Unbiasedness for SATT

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.

$$SATT = \frac{1}{N_1}\sum_{i:D_i=1} Y_{1i} - \frac{1}{N_0}\sum_{i:D_i=0} Y_{0i} \tag{15}$$

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

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1} Y_{0i} - \sum_{D_i=0} w_i Y_{0i} \tag{16}$$

$$= \frac{1}{N_1}\sum_{i:D_i=1}\left(\phi(X_i)^\top\theta + \epsilon_i\right) - \sum_{i:D_i=0} w_i\left(\phi(X_i)^\top\theta + \epsilon_i\right) \tag{17}$$

$$= \theta^\top \frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i) + \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \theta^\top\sum_{i:D_i=0} w_i\phi(X_i) - \sum_{i:D_i=0} w_i\epsilon_i \tag{18}$$

$$= \theta^\top\left(\frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i) - \sum_{i:D_i=0} w_i\phi(X_i)\right) + \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i \tag{19}$$

$$= 0 + \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i \tag{20}$$

The bias is the expectation of this quantity,

$$bias = \mathbb{E}\left[\widehat{DIM}_w - SATT\right] \tag{21}$$

$$= \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i\right] = 0 \tag{22}$$

A.2.1 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 the 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

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i + \frac{1}{N_1}\sum_{i:D_i=1} h(X_i) - \sum_{i:D_i=0} w_i h(X_i) \tag{23}$$

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.

A.3 Illustration of worst-case bound on bias

Next, a simulation is useful to illustrate the effectiveness of the worst-case bound. This involves five steps:

1. Randomly draw a function from the space of functions corresponding to a Gaussian kernel: first choose a set of 100 "knots", $Z_j \sim Unif(0,1)$ for $j \in 1,...,100$, then choose a kernel function $k(\cdot,\cdot)$ (Gaussian with $b = 0.1$), and randomly draw the c vector according to $c_i \sim N(0,1)$ for $i \in 1,...,100$. Together these quantities characterize a chosen function in the RKHS of kernel k, $f(\cdot) = \sum_j c_j k(\cdot, Z_j)$.

2. For $N = 25$ and $N = 200$ (see below), randomly draw covariate data $X_i \sim Unif(0,1)$. Generate the values of Y_{0i} under the function drawn above, i.e. $Y_{0i} = \sum_j k(X_i, Z_i)$.

3. Draw treatment status values, D_i, with the probability of treatment relating to the values of $Y_i(0)$. Specifically let $\tilde{y}$ be the centered and standardized $Y_i(0)$, then let $D_i = 1$ with probability $\frac{1}{1+exp(-\tilde{y})}$ and 0 otherwise.

4. Choose treatment effect size, $TE = 1$. Construct the observed outcome, $Y_i = Y(0) + D * TE$.

5. The proposed method is then used to estimate the treatment effect. The estimated treatment effect minus the true one gives the bias due only to incomplete balance, as there is no noise added to the outcome and treatment effects are constant here.

This whole procedure is iterated 1000 times and the bound is computed for each iteration, with an estimate of the Hilbert norm (γ) obtained by using regularized kernel regression with the same kernel as used for kernel balancing (using the `KRLS` package for `R`).[7] Figure 5 shows the result, with N of just 25, and then with N of 200. Only with extremely small sample sizes (i.e. $N = 25$) does this bias come near the bound, and no bias was greater than the bound would allow. As the sample size grows, the bound becomes increasingly conservative. This is to be expected: the bias hits the bound only when the coefficients in the eigenvector space (d) happen to be perfectly aligned with the imbalances on the eigenvectors. Such an alignment becomes increasingly unlikely as the sample size, and thus number of eigenvectors, grows.[8]

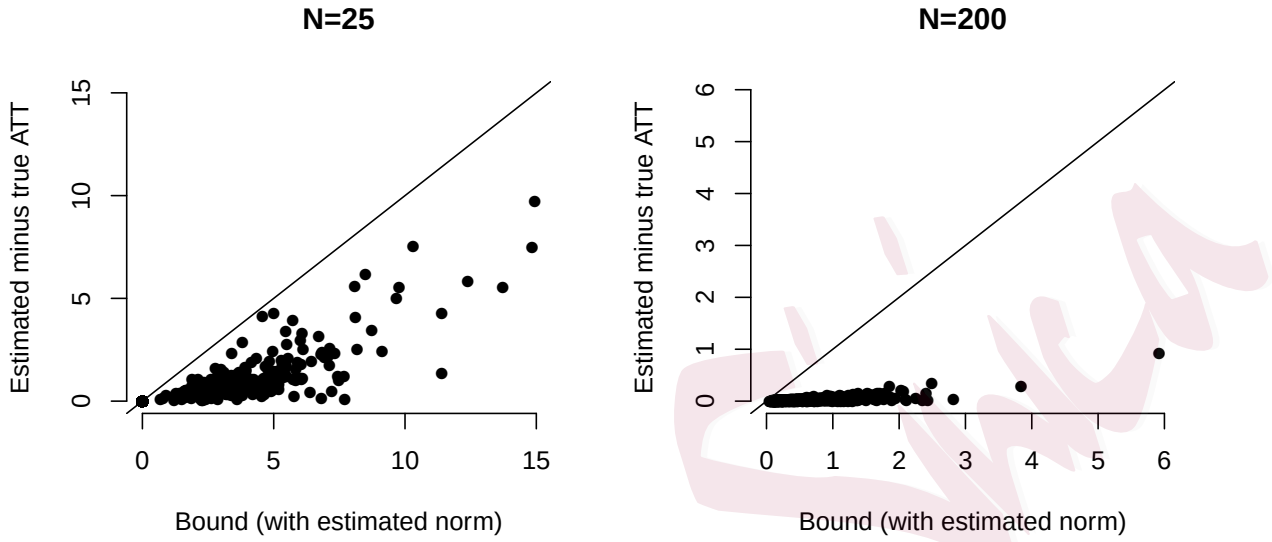
A.4 Balance on an approximation of K

The main text focuses on minimizing the worst-case bias due to imperfect balancing as the rationale for achieving balance on eigenvectors with larger eigenvalues. A closely related justification begins by seeking a lower-dimensional approximation of $\mathbf{K}$ that is most similar in some respect.

[7] The `KRLS` package uses generalized (leave-one-out) cross-validation to choose the regularization parameter, λ. This in turn determines the Hilbert space norm of the selected regression function, which can be computed using the empirical value of $\hat{c}^\top \mathbf{K} \hat{c}$, providing a reasonable choice of γ that corresponds well to the observed data.

[8] Future work could fruitfully derive less extreme values, such as the expectation of this bias. This however would likely require further assumptions over the probability distribution of functions in the ball of the Reproducing Kernel Hilbert Space carved out by γ.

Figure 5: Simulation Illustrating Behavior of Bias Bound



The worst-case bound on the error in the ATT estimate due to balancing on selected eigenvectors $\mathbf{K}$ rather than the entire matrix. Each of 200 iterations draws a new function at random from the RKHS, a new set of data at which to evaluate it, and a new estimate of the ATT using the method proposed here. The horizontal axis shows the estimated bound on this error from each simulation. The vertical axis shows the actual error. The required Hilbert norm, γ, is estimated using `KRLS`, in which leave-one-out cross validation is used to determine the appropriate complexity of the function. *Left:* All errors are less than the estimated bound, however with the very small sample size of 25, some errors come near the bound. *Right:* With $N = 200$, the bound on the error becomes highly conservative.

Suppose we have rank-r approximation to $\mathbf{K}$, $\widetilde{\mathbf{K}}^{(r)}$. We might seek the $\widetilde{\mathbf{K}}^{(r)}$ closest to $\mathbf{K}$ in the Frobenius norm, i.e. minimizing

$$||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_{\mathcal{F}} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} |\mathbf{K}_{i,j} - \widetilde{\mathbf{K}}^{(r)}_{i,j}|^2}$$

. Alternatively, and closely related to *biasbound*, recall that our aim in achieving mean balance on $\mathbf{K}$ is to ensure that any linear projection $\mathbf{K}c$ for some $N \times 1$ vector c has equal means in the two groups. In choosing a rank r approximation, we thus want to ensure that for c of a particular size $||c||$, $\widetilde{\mathbf{K}}^{(r)}c$ and $\mathbf{K}c$ are as close as possible. Thus, it is desirable to minimize the operator 2-norm, $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_2 = sup\ \frac{||\mathbf{K}c - \widetilde{\mathbf{K}}^{(r)}c||_2}{||c||_2}$. Among all rank r matrices, the choice of $\widetilde{\mathbf{K}}^{(r)}$ minimizing both $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_2$ and $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_{\mathcal{F}}$ is given by principal components analysis (PCA; Eckart and Young, 1936). Since PCA constructs $\widetilde{\mathbf{K}}^{(r)}$ as a linear projection of the first r eigenvalues $\mathbf{K}$, we need not actually work with the projected approximation $\widetilde{\mathbf{K}}^{(r)}$ – we can simply work directly with the eigenvectors themselves. That is, the eigenvectors of $\mathbf{K}$ (which we obtain here using singular value decomposition rather than PCA here) provides a new set of bases, and we will seek balance on the first r of them (as ordered by the corresponding eigenvalues). This provides an N by r matrix of orthonormal bases for which we attempt to make the control group have the same mean for the treated by weighting.

A.5 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:

$$\begin{aligned}
\mathbb{E}[Y_{0i}|D=1] &= \mathbb{E}_x\left[\mathbb{E}[Y_{0i}|X, D=1]\right] && (24)\\
&= \theta^\top \int \phi(x) p(x|D=1)dx && (25)\\
&= \theta^\top \mathbb{E}[\phi(x)|D=1] && (26)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_w[Y_{0i}|D=0] &= \mathbb{E}_{w,x}\left[\mathbb{E}[Y_{0i}|X, D=0]\right] && (27)\\
&= \theta^\top \int \phi(x) w p(x|D=0)dx && (28)\\
&= \theta^\top \mathbb{E}_w[\phi(x)|D=0] && (29)
\end{aligned}$$

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.

A.6 Proof of proposition 1

Proposition 1 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 b, 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 $\frac{1}{N_1\sqrt{2\pi b}} K_t \mathbf{1}_{N_1}$ 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, $\frac{1}{N_0\sqrt{2\pi b}} K_c \mathbf{1}_{N_0}$ 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 $\frac{1}{N_0\sqrt{2\pi b}}$ 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 $\frac{1}{\sqrt{2\pi b}} K_c w$, for vector of weights w. If weights are selected so that this equals the density of the treated:

$$\begin{aligned}
\frac{1}{N_1\sqrt{2\pi b}} \mathbf{K_t} \mathbf{1}_{\{N_1\}} &= \frac{1}{\sqrt{2\pi b}} \mathbf{K_c} w\\
\frac{1}{N_1} \mathbf{K_t} \mathbf{1}_{\{N_1\}} &= \mathbf{K_c} w\\
\overline{K_t} &= \mathbf{K_c} w\\
\overline{K_t} &= \overline{K_c(w)} && (30)
\end{aligned}$$

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.

A.7 Derivation of $\phi(X_i)$ for Gaussian Kernel

While the functions linear in $\phi(X_i)$ corresponding to a Gaussian kernel can more easily be understood as those that can be formed by superposing Gaussian kernels over the observations, one may also explicitly construct features $\phi(X_i)$ consistent with the requirement that $K(X_i, X_j) = \langle \phi(X_i).\phi(X_j) \rangle$ for the standard inner-product. One simple approach is, setting $b = .5$ for convenience, yields:

$$k(X_i, X_j) = exp(-||X_i - X_j||^2/1) \tag{31}$$
$$= exp(-X_i^2)exp(-X_j^2)exp(2X_i, X_j) \tag{32}$$
$$= exp(-X_i^2)exp(-X_j^2)\sum_{d=0}^{\infty} \frac{2^d X_i^d X_j^d}{d!} \tag{33}$$

where the last line follows by a Taylor series expansion of $exp(2X_iX_j)$. Finally the division of terms can be completed, as:

$$k(X_i, X_j) = \sum_{d=0}^{\infty} \sqrt{\frac{2^d}{d!}} exp(-X_i^2 X_i^d) \sqrt{\frac{2^d}{d!}} exp(-X_i^2 X_i^d) \tag{34}$$

This is simply an inner product of two infinite-dimensional vectors of the form

$$\phi(X_i) = \left[\sqrt{\frac{2^0}{0!}} exp(-X_i^2 X_i^0), \ \sqrt{\frac{2^1}{1!}} exp(-X_i^2 X_i^1), \ \ ..., \ \ \sqrt{\frac{2^\infty}{\infty!}} exp(-X_i^2 X_i^\infty) \right] \tag{35}$$

Figure 6 considers a one dimensional covariate, X, and shows what value each of the first 5 of these features would have at various values of X.

A.7.1 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 7 shows the resulting density plots, using density estimates provided by `kbal` in which the density of the treated is given by $\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t 1}_{N_1}$ and the density of the controls is given by $\frac{1}{N_0\sqrt{2\pi b}}\mathbf{K_c 1}_{N_0}$. 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.

A.8 Inverse Propensity Score Weights as Multivariate Density Equalization

It is useful to show more explicitly the role played by inverse propensity score weights in estimating the ATT, as this leads to an appreciation of how these weights relate to multivariate density equalization, and the sense in which they are equivalent to the kernel balancing weights despite flowing from different initial goals.

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

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=1] && (36)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int \mathbb{E}[Y_{0i}|D_i=1,x]p(x|D_i=1)dx && (37)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int \mathbb{E}[Y_{0i}|D_i=0,x]p(x|D_i=1)dx && (38)\end{aligned}$$

Expression 38 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,

$$\begin{aligned} \text{DIM} &= \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0] && (39)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int \mathbb{E}[Y_{0i}|D_i=1,x]p(x|D_i=0)dx && (40)\end{aligned}$$

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,

$$\begin{aligned} \text{DIM}_w &= \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}_w[Y_{0i}|D_i=0] && (41)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int w_i\mathbb{E}[Y_{0i}|D_i=1,x]p(x|D_i=0)dx && (42)\end{aligned}$$

where w_i is a function of X that allows us to upweight or downweight control units. The difference between expression 38 and 40 can be resolved by choosing weights on the control units,

$$w_i = \frac{p(x|D_i=1)}{p(x|D_i=0)} \tag{43}$$

Through Bayes theorem, we can replace the class densities in this expression with more familiar propensity scores to obtain $w_i = \frac{p(D_i=1|x)p(D_i=0)}{p(D_i=0)|x)p(D_i=1)}$. Since $D_i = 0$ for all units given weights, 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. If properly estimated, these would ensure that the whole distribution

of X for the control units is adjusted to equal the distribution among the treated.

Note that in the form 43, it becomes clear that were we to adjust the sample to make treated and control groups have the same distribution of covariates, these weights would become constant and thus unnecessary. This is achieved, insofar as the smoothed multivariate densities on which kernel balancing obtains balance are reasonable approximations of the true densities. In this sense, kernel balancing achieves the goals of inverse propensity score weighting, but has the advantage of avoiding any functional form assumption or direct estimation of the propensity score.

A.9 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 interpretation given above. Specifically, the density estimators at all points is constructed using the kernel matrix. Then, treated units are trimmed if $\frac{p_{X|D=1}(x_i)}{p_{X|D=0}(x_i)}$ 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.

A.10 Detailed Choice of Kernel

Using the kernel as defined by 7 for some choice of b, any continuous function $\mathbb{E}[Y_{0i}|X_i]$ can be consistently estimated by functions linear in $\phi(X_i)$. However, some kernel choices work better than others in a sample of limited size. Accordingly, in machine learning applications utilizing kernels, it is common to consider details of the kernel definition that may improve the ability to fit the target function linearly in $\phi(X_i)$ (or equivalently, the columns of $\mathbf{K}$) when the sample size is limited.

The first consideration of this type is how X is scaled and rotated. If some variables in X_i have variances orders of magnitude larger than others, the columns of $\mathbf{K}$ will reflect mostly distances on the highest-variance variables, providing little information on distances among the smaller variables. This is unproblematic as the sample size grows to infinity – the superposition of Gaussians will still allow flexible modeling of the target functions in the limit. But in a small sample, it limits the quality of fit. It is thus common to utilize a Gaussian kernel that computes the Euclidean distance over variables that have been rescaled to have the same variance. This also has the benefit of making the

results invariant to any unit-of-measure decisions. Kernel balancing utilizes this approach. Beyond this, some investigators also wish to make the results invariant to rotation, utilizing a Mahalanobis distance rather than Euclidean distance in the Gaussian kernel. This is left as an option in kernel balancing as implemented here.

Second, b must be chosen. Since mean balance on Y_{0i} is the primary goal, not density estimation or equalization, the choice of the kernel and b should be made accordingly. While it is tempting to think of b as the usual bandwidth that must be carefully selected in density estimation procedures, here the choice of parameter b is understood first as 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. The choice of b also has implications in terms of a bias-variance tradeoff and feasibility: if b is too large, mean balance is easier to achieve and the weights will typically have lower variance, the resulting balance is less precise (and the corresponding smoothed densities more "blurred"). At the extreme as b approaches infinity, $\mathbf{K}$ approaches a matrix with 1 in every position, indicating that all observations are alike and nothing needs to be done to obtain balance. In the opposite extreme, as b becomes very small, $\mathbf{K}$ begins to approximate the identity matrix. 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 b is discussed in Appendix A.9). The interesting cases lie in between these extremes, where choices can be made to "blur" the features more or less in order to make balance easier to achieve on more dimensions of $\mathbf{K}$. Note that standard matching and weighting methods typically involve a bias-variance tradeoff as well, though it may be implicit or difficult to manipulate directly. For example, in matching, the number of control units matched to each treated unit, as well as the choice of caliper, and of course the choice of how many covariates to match on all have implications for the bias-variance tradeoff. In Coarsened Exact Matching Iacus et al. (2011), the size of the bins used to coarsen each covariates have direct bias-variance implications. King et al., 2017 usefully discusses the related "balance vs. sample-size frontier". Kallus (2016) discusses the bias-variance tradeoff in optimal matching procedures and the assumptions under which a mean squared error criterion is minimized by various procedures. In related weighting methods such as Hainmueller (2012), there is no direct control of the bias-variance tradeoff except implicitly through the set of covariates one is seeking (exact) mean balance on[9] Likewise, if one uses a propensity score model to choose inverse propensity score weights, for example, the bias-variance implications of those models are difficult to control.

Because there is no "right answer" as to what b should be, I provide here three guidelines for transparent reporting. First, a useful reporting standard would be to provide results at $b = dim(X)$, while also showing results at other choices for robustness. The reason to choose $b = dim(X)$ (the default value used above) is that the square of $\mathbb{E}[||X_i - X_j||]$, used in the exponent of the kernel calculation (7) scales with $dim(X)$. Choosing b 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 remains open to debate, but the choice of $b = dim(X)$ has offered good performance in most cases (though higher values tend to perform more reliably when balance is very difficult to achieve, as in the National Supported Work example). Second, the degree to which weights become large and uneven should be reported. I propose the quantity *min90*, which is the minimum number

[9]In principle, setting the tolerance or stopping point for convergenge of the algorithm, or other procedures, could be added to allow a measure of control.

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 National Supported Work example reported above demonstrates this. Third, investigators may wish to present their results across a range of b values to ensure this choice is not consequential in a given application (see King et al., 2017 for a related proposition regarding matching estimators and the "balance/sample-size" tradeoff). Should the results vary across b values, inspecting L_1 and the concentration of weights (e.g. through $min90$) can be helpful for understanding the bias-variance consequences of a given choice.

Fortunately, the choice of b is easy to address in many cases because a wide range of b values often allow large improvements in L_1 paired with stable ATT estimates. Following the recommendation above to show estimates at $b = 1p$ and across a rage of b values, Appendix A.11 shows ATT estimates for kernel balancing with the standard covariate set in the National Supported Work empirical benchmark (Section 4). While there is some variation in estimates when b is small ($2p$ or less), the estimates stabilize above that to values of over $50p$. Moreover, despite the potential for a bias-variance tradeoff, when good balance can be achieved on even the smaller b values without resorting to extreme weights, then the variability of ATT estimates (i.e. under resampling) can remain stable across a wide range of b values. Appendix A.11 also shows boxplots of ATT estimates for the simulated example (Section 3.1), showing both low bias and stable variance across the range of attempted b values.

A.11 Stability Across b in simulation and empirical example

As described in the text, kernel balancing is the only method of those attempted that approaches unbiasedness in estimating the simulated effect of peacekeeping when a non-linear function of the covariates was confounding. The only parameter that must be chosen by the user is b, though `kbal` provides a default of $b = dim(X)$. In Figure 8, we see that this result is largely insensitive to the choice of b ranging from one-quarter to four times the default. If anything, ATT estimates improve with b somewhat above the default, though setting b larger can come at the cost of more extreme weights in some natural datasets where overlap in the covariate distributions may not be as good.

We also examine sensitivity of results to the choice of b in the National Supported Work benchmark example. As described in the text, kernel balancing is used in three ways: with the original set of 10 covariates used previously (*standard*), with a set of covariates that replaces the income variables with their logs (*log*), and with a set that, for the three continuous variables, includes their squares (*squares*). At the defaults values of b (the number of covariates), each estimate does well as shown in the text. Figure 9 shows the results for all three specifications over a wide range of b values. While there is some instability at low levels of b – due to the difficult of achieving balance in this example – past that point, the ATT estimates at the chosen weights are extremely stable and accurate. The guidelines suggested here is that investigators routinely report their estimates across a range of b estimates to avoid selective reporting of results.

A.12 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 10, 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 11 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 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.

A.13 Are democracies more selective?

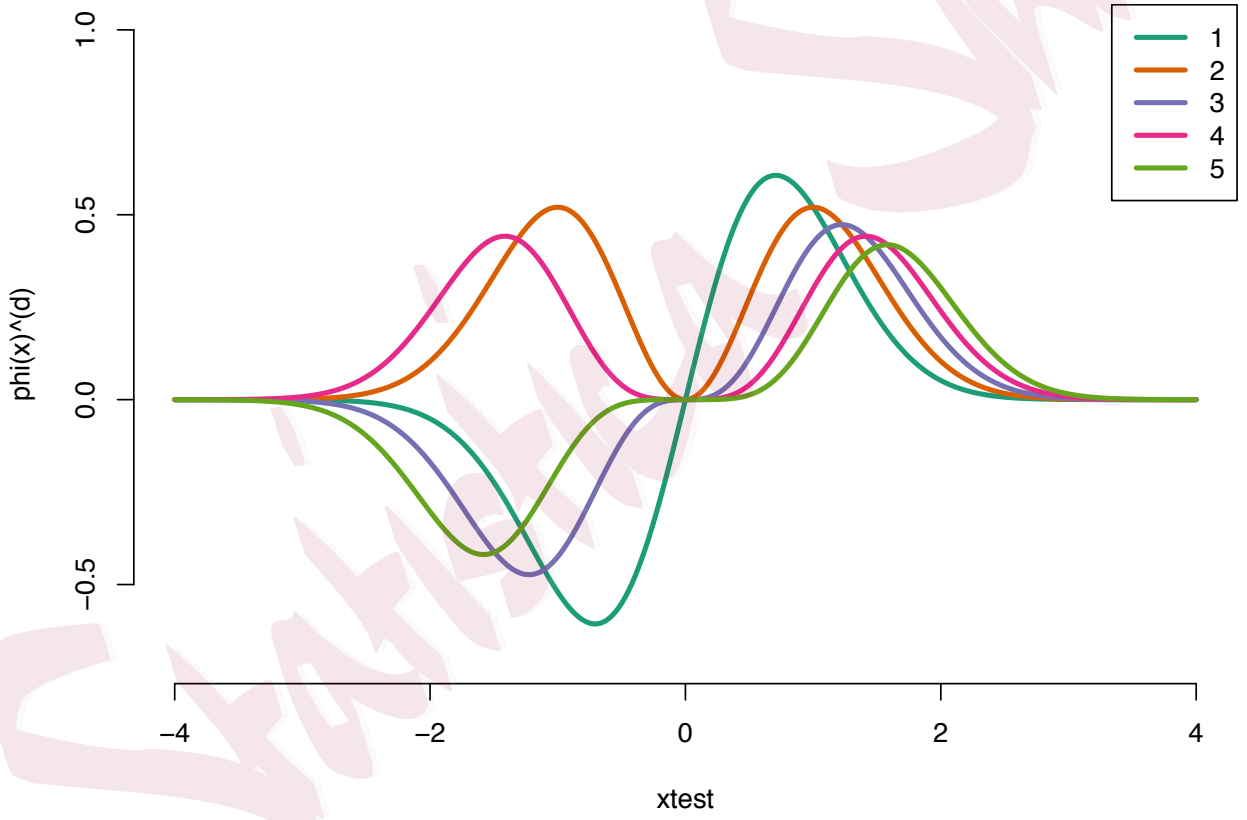
One puzzle regarding the claim that democracies are inferior counterinsurgents has been why democracies, whatever their weaknesses as counerinsurgents, are not also better able to "select into" conflicts they are more likely to win. The same qualities that are theorized to make democracies more susceptible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

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,

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

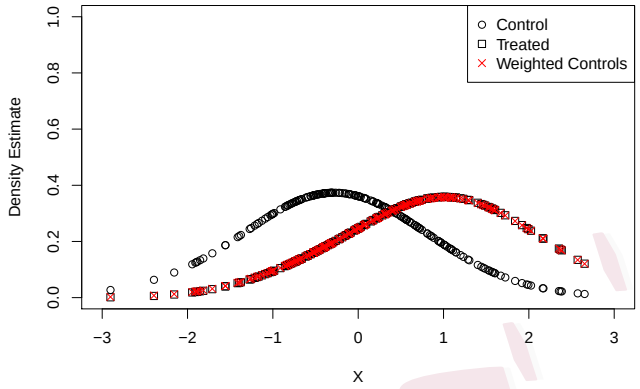
That is, the naive difference in means is the average treatment effect on the treated (had they fought in the same types of cases), plus a selection effect indicating how democracies and non-democracies differ in their probabilities of victory based only on fighting different types of cases (i.e. in 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.

Figure 6: First five values of $\phi(X)$ at varying values of X



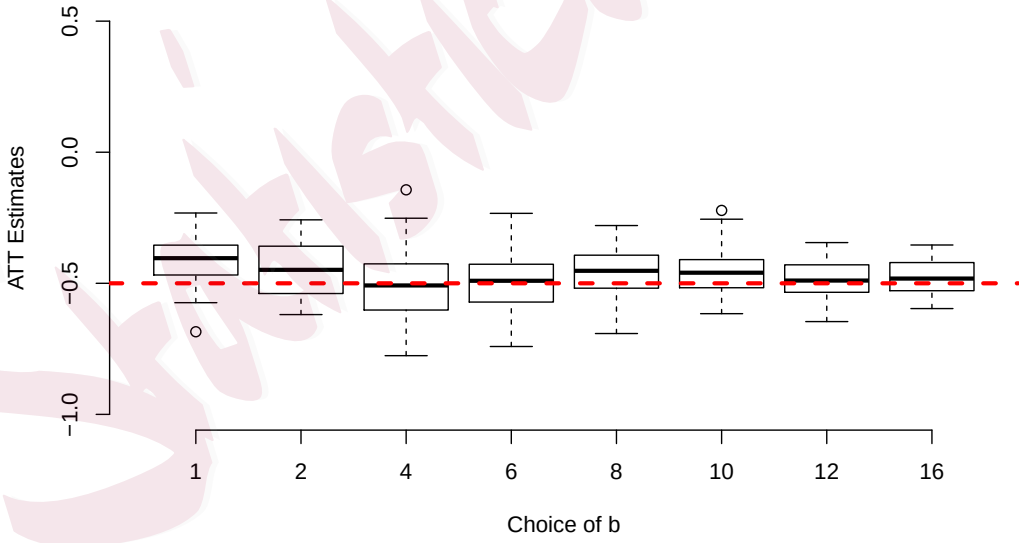
Explicit view of $\phi(X_i)$ for one choice of $\phi(X_i)$ consistent with $K(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$ for a Gaussian kernel K as described in Equation A.7.

Figure 7: Density-Equalizing Property of Kernel Balancing



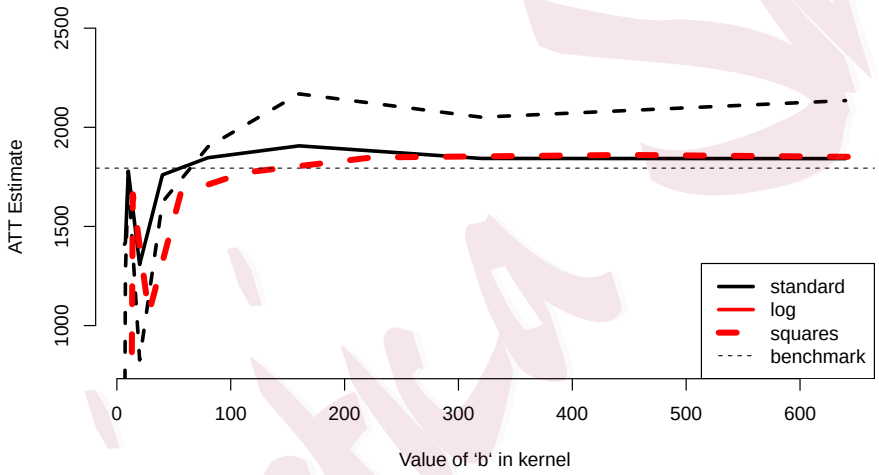
Plot showing the density-equalization property of kernel balancing. For 200 observations of $X \sim N(0,1)$, treatment is assigned according to $Pr(treatment) = 1/(1+exp(2-2X))$, producing approximately two control units for each treated unit. Black squares indicate the density of the treated, as evaluated at each observation's location in the dataset (and given the choice of kernel and b). 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.

Figure 8: Simulation: sensitivity to choice of p



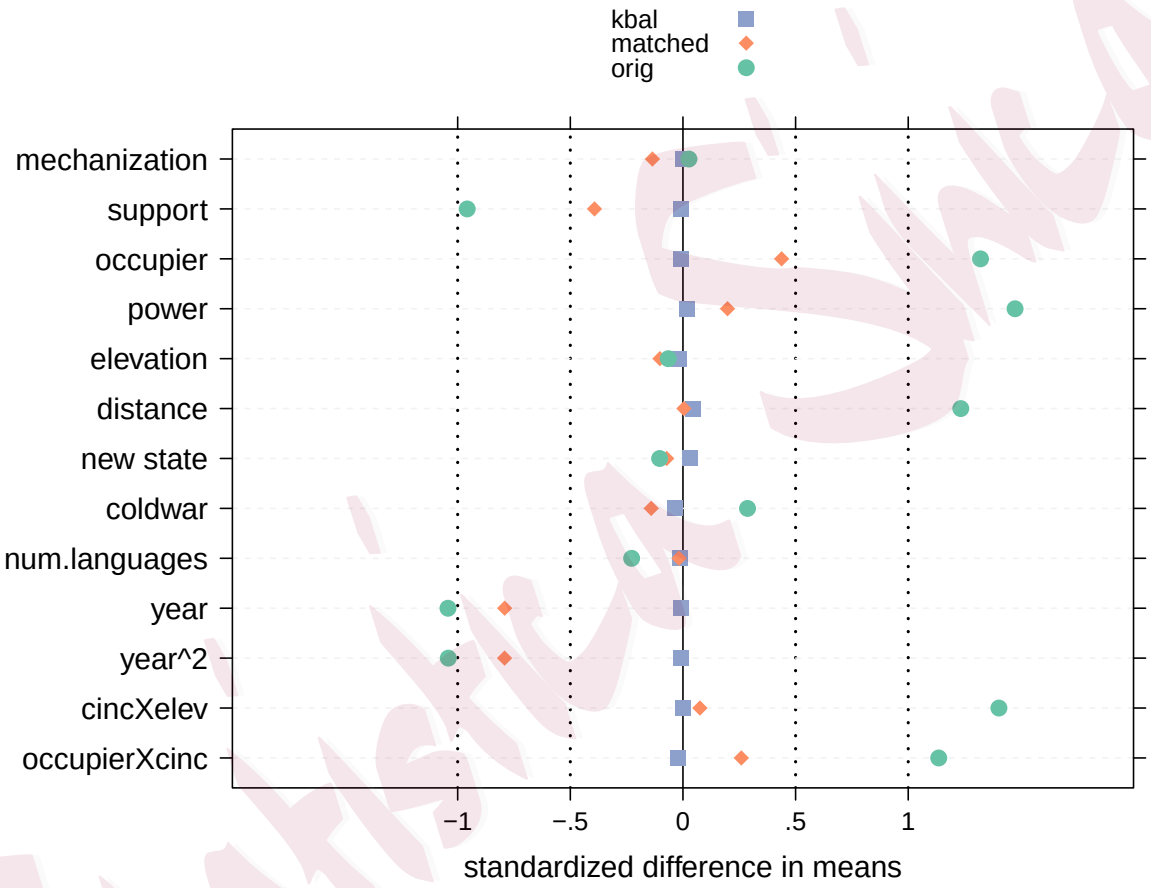
Boxplot illustrating distribution of average treatment effect on the treated (ATT) estimates using *kernel balancing*, as the bandwith parameter b is varied. At each value of b, 50 simulations are used, each drawing a separate dataset from the same data generating process. The default choice of b is $dim(X) = 4$, with results here shown from one-quarter to four times that value. The actual (population) ATT is -0.5, indicated by the dashed line. Results show very low bias at all values of b.

Figure 9: Stability of National Supported Work estimates to choice of b



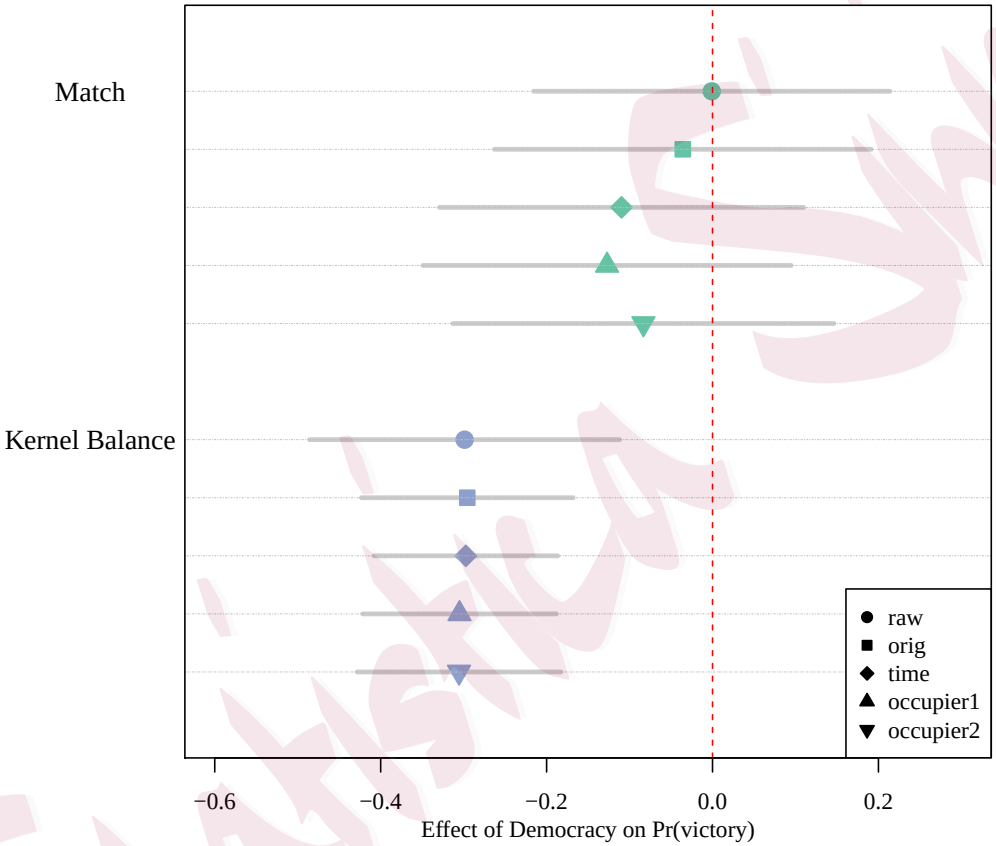
Stability of ATT estimates for *kbal* across different choices of b in the National Supported Work Example. The *standard* estimates employed all 10 of the covariates provides in the Dehejia & Wahba dataset Dehejia and Wahba (1999) and typically used for this task. The *log* specification makes the choice to replace the income variables with their logs (plus one), and the *squares* model add squared terms for the continuous variables. While there is some instability at small choices of b, the results are remarkable stable and close to the benchmark at higher values of b.

Figure 10: 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.

Figure 11: 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.