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

Abstract

Matching and weighting methods are widely used to estimate causal effects when unobserved confounders are assumed not to exist. Matching is appealing for its non-parametric nature, but with continuous variables, is not gauranteed to fully condition on the observables or remove bias due to observables. Weighting techniques can be used to make pre-specified functions of the covariates have exactly equal sample means for the treated and control group. However, these weights assure unbiased effect estimation only when the potential outcomes are assumed to be linear in those pre-specified functions of the observables. Kernel balancing is a weighting technique that relaxes this limitation by choosing weights such that the treated and control group have equal means on a set of bases implied by a kernel. This ensures unbiasedness of the difference in (weighted) means estimator so long as the potential outcomes are assumed to be linear in these kernel bases, which can be proven to be a very large space of continuous, non-linear, non-additive functions. I describe in detail what this space of functions looks like in the case of the Gaussian kernel. In practical terms, kernel balancing offers a reasoned solution to the long-standing question of which functions of the covariates investigators should match, weight, or check balance on, reducing researcher degrees of freedom and avoiding an iterative process of matching or weighting in different ways. I further show a duality by which the weights chosen in this way are also the weights that make the estimated empirical density of covariates the same for the treated and control – when the same choice of kernel is used to make those density estimates. The approach is fully automated up to the choice of a kernel parameter, for which I provide a default option. An `R` package, `KBAL`, implements this approach.

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

1 Introduction

In seeking to make causal inferences from observational data under the assumption of no unobserved confounding, the goal of matching and weighting techniques is to make the distribution of observables the same for the treated and control group, as a means of effectively "controlling for" or "adjusting for" differences on the observables. However, when exact matching is not possible – such as when continuous variables are included – the distribution of observables cannot be made exactly the same in the two groups.

A simple example, explored at greater length in Section 3, illustrates the problem to be solved. 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 drives the probability of receiving treatment. Further suppose that both the probability of receiving the treatment and the (potential) outcomes are 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: among treated and control units paired together by nearest neighbor matching (for example), the treated unit is on average more likely to be higher on X_1/X_2. These non-random matching discrepancies do 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.

Weighting approaches exchange this problem for another, getting "exact balance" on desired moments, but sacrificing the non-parametric quality of matching. Consider a weighting estimator that chooses weights to obtain equal means on X_1 among the treated and control, and likewise on X_2. Though it is often possible to achieve exactly equal means on these two covariates (referred to here as "mean balance"), this does not in general imply that X_1/X_2 has equal means for the two groups. This is problematic: X_1/X_2 influences the potential outcomes, so unequal means in the treated and control groups will lead to unequal mean potential outcomes for the treated and control 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 fail to make the treated and control groups "comparable" enough on even the observables to complete the desired adjustment for observable differences. If the investigator happened to know that X_1/X_2 is critical to both the treatment assignment and potential outcomes this could be avoided. However, my supposition is that investigators rarely if ever have sufficient theoretical knowledge to unfailingly guess these functional forms. Moreover, allowing the investigator the leverage to guess at such functional forms creates the opportunities for selective reporting of results.

The simulations in Section 3 furhter examines a similar hypothetical example, showing how severely even simple non-linear functions of the observables such as this can generating large biases from state-of-the-art matching and weighting estimators. Kernel balancing mitigates this problem, achieving nearly equal means on X_1/X_2 without the investigtor 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, as with matching, regression, and other estimation procedures, we must assume that all confounders are observed. Second, we assume that the regression surface for the non-treatment potential outcome (Y_{0i}) falls in the (Reproducing Kernel Hilbert) space associated with a choice of kernel. Here, I propose using a Gaussian kernel, as the resulting function space is suitable to a wide variety of smoothly varying outcomes. Third, despite the complexity of this space of outcome models, we find that models in this space can be represented as those linear in a (potentially infinite dimensional) basis expansion, $\phi(X_i)$.

While that proves mathematically very useful, more practically, this space of functions is representable as those linear in the N-dimensional rows K_i of a kernel matrix, as in $\mathbb{E}[Y_{0i}|X_i] = K_i^\top c$. Fourth, we can choose weights on the control units such that the weighted average K_i among the controls equals that average K_i among the treated. Because the regression surface for Y_{0i} is linear in K_i, equal means on K_i automatically ensures equal means on Y_{0i} for the treated and weighted control groups. This holds without having to estimate the coefficients (c) or otherwise fitting any model. Finally, a simple difference in means estimator (with the corresponding weights) proves unbiased for the ATT due to this equal means on potential outcomes. The analytical approach below, while more detailed and rigorous, flows from this logic.

Kernel balancing does not directly seek to make the multivariate distributions of the covariates equal for the treated and control groups, as matching and weighting techniques generally seek to approximate. Rather, it relies on the weaker estimation goal of achieving equal means on Y_{0i} in the two groups. Nevertheless, it has an illuminating interpretation in terms of multivariate distributions of the covariates: the weights chosen by kernel balancing are also those that equalize the *estimated* empirical multivariate distributions of covariates for the treated and control, *when estimated using the same choice kernel* as a smoother. This provides an elegant link between the assumption one is willing to make about the space of outcome models (e.g. that is well modeled by a Gaussian kernel) and the choice of smoother for which the estimated empirical distribution of covariates are made equal. It also gives intuition for these weights as those that "move" the density of control observations in the covariate space to be equal to that of the treated, up to the precision of the associated density estimator.

To briefly place kernel balancing in context of other methods, compared to approaches that depend on fitting outcome models (such as regression), kernel balancing still relies on an assumed outcome model space, but no outcome model is ever fitted. Moreover, 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 heroic modeling assumptions to bridge potentially large gaps between the location of control and treated observations. However, unlike matching, kernel balancing avoids iterative matching and balance-checking procedures, and more importantly, avoids the bias due to inexact matches. Relative to propensity score approaches Rosenbaum and Rubin (1983b), kernel balancing requires no model for the propensity score and thus avoids the severe biases (see e.g., Smith and Todd, 2005; Kang and Schafer, 2007) due to potential misspecification of the propensity score. Finally, the method is most similar to other weighting procedures that use only covariate information (such as Hainmueller, 2012) and related survey weighting procedures going back at least to Deming and Stephan (1940).

In this context, a practical contribution of kernel balancing is that it provides one principled (and automated) answer to the question of what functions of the covariates must have equal means in order to produce an unbiasedness result, given an assumed but flexible, smooth space of outcome 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 smooth functions of those covariates. Much more discussion follows below regarding how kernel balancing relates to existing procedures both in theory and in terms of its performance on simulated and real data applications.

In what follows, Section 2 provides the analytical framework and develops the method. Section 3 provides a basic simulation, highlighting the dangers ineherent 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 effectivness in recovering an experimental benchmark from observa-

tional data, using the National Supported Work demontration (LaLonde, 1986). Section 5 discusses the implications of this procedure, further details (including important feasibility and approximation issues), and further comparison to existing matching, weighting, regression, and propensity score approaches. It also provides intuitions for the nature of assumptions made on the space of models for the potential outcomes and why they are suitable for a broad class of problems. Following conclusions in Section 6, additional proofs, discussions, simulations, and empirical examples can be found in the 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. Using the Neyman-Rubin potential outcomes framework (see e.g. Rubin, 1990; 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 $\mathcal{X}$ is the support, assumed to lie in $\mathbb{R}^P$. These are assumed to be unaffected by the treatment and are thus called "pre-treatment" covariates. 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 $p(X, Y_1, Y_0, D)$.

A critical assumption required throughout is that treatment assignment is ignorable with respect to the potential outcomes, conditionally on the covariates. I will refer to this as "conditional ignorability", to emphasize that the treatment assignment is assumed ignorable *only conditionally on the covariates*, though the same assumption is sometimes called "strong ignorability (conditional on covariates)" (Rosenbaum and Rubin, 1983a), "no unobserved confounding", or "selection on observables".

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.

At it's core, 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 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 N, is an expanded set of these characteristicss to be used as a set of basis functions. 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 $\phi(x)$:

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

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

Note that a similar assumption can be made regarding $\mathbb{E}[Y_{1i}|X_i]$, and is required for analysis of the ATE or ATC, but not the ATT, which I focus on here for ease of exposition.

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).

2.2 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 a simple estimand for this, the difference in means,

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

which replaces the unobservable second term in the ATT expression ($\mathbb{E}[Y_{0i}|D_i = 1]$) with observable 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 in turn implies:

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

Both Expressions 2 and 3 have very useful direct interpretations. The equality in 2 can be interpreted as requiring $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$ for unbiasedness of the DIM. This has a natural intuition: the expected non-treatment outcome among the treated is unobservable, and the DIM seeks to replace it with the expectation of the non-treatment outcome among the controls, which is observable. But this substitution only works if the two are equal. However, it is Equation 3 which gives us a feasible 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.

The core idea of kernel balancing is that, having chosen suitable $\phi(\cdot)$ as bases, we need only obtain "equal means" on $\phi(X_i)$ rather than on the original X_i. Let us refer to this as "mean balance on $\phi(X_i)$". In actuality this will not hold in the natural data, but rather will be the objective of a weighting procedure (see next). In Section 2.5, I describe how weak Assumption 2 can be made by using kernels to choose $\phi(\cdot)$.

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

Our aim is to make mean balance on $\phi(X_i)$ (Equation 3) hold through a weighting procedure when it does not already hold. To avoid confusion, note that $p(x|D_i = 1)$ does not generally equal $p(x|D_i = 0)$ naturally in the applications of interest here. Rather, even under Assumption 1, the distribution of X_i among the treated and control may differ, giving rise to the different expected potential outcomes in the treated and control group, which we seek to address by obtaining mean balance on $\phi(X_i)$.

On the population level, consider an adjustment procedure by considering a function of the covariates $\tilde{g}(X_i)$, with the property that:

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

Where $g(X_i) = \tilde{g}(x)p(x|D_i = 0)$ is scaled to integrate to 1 as a new density, which can be used to construct an effectively "g-weighted" expectation of $\phi(X_i)$ among the controls. Written as $\tilde{g}(x)p(x|D_i = 0)$, we see that $\tilde{g}(x)$ reweights the natural distribution of X_i among the controls. Setting $\tilde{g}(x) = \frac{p(x|D_i=1)}{p(x|D_i=0)}$ would be one choice that satisfies this, directly making $g(x) = p(x|D_i = 1)$ (see Section 5 for equivalence to inverse propensity score weighting). Critically, 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 DIM estimand (1) of the ATT can now be modified by these weights to become simply

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

which uses the weights implied by $g(x)$ in order to ensure that the expectation of $\phi(X_i)$ taken over the control population equals $\mathbb{E}[\phi(X_i)]$ over the treated population, thus making $\mathbb{E}[Y_{0i}]$ equal for the two groups.

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, makes the expectation of the non-treatment potential outcome for the untreated equal to that of the treated, i.e. $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_g[Y_{0i}|D_i = 0]$. And that condition in turn ensures that the DIM estimand equals the ATT.

2.4 Sample DIM and Weights

We now turn to the sample and corresponding choice of weights. In Equation 4, the expectation of $\phi(X)$ among the treated, $\mathbb{E}[\phi(X_i)|D_i = 1]$, is replaced by its sample analog, the sample mean $\frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i)$. 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 corresponding to the population contraints in Equation 4. This replaces $\sum_{i:D_i=0}\phi(X_i)w_i$, 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,

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

This brings us to the main result,

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

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

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

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* $p(X, Y_1, Y_0, D)$*.*

This derives simply from replacing each expectation in 5 by the corresponding sample average and weights with the corresponding sample moment conditions. Alternatively, for proof one may also begin with a given sample, showing that in 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. Unbiasedness of this sample estimator for the SATT is proven in Appendix A.2. That approach also allows 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 $\phi(\cdot)$

Thus far, we have developed a framework for estimation of the ATT by weighting that clarifies the need for: conditional ignorability (Assumption 1), a set of bases $\phi(X_i)$ in which $\mathbb{E}[Y_{0i}|X_i]$ is assumed to be linear (Assumption 2), and weights that obtain expectation of $\phi(X_i)$ among the treated and controls. A wide range of basis expansions $\phi(\cdot)$ could in principal be chosen under this estimation framework. Here, I propose a strategy of not chosing $\phi(\cdot)$ 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.

The fundamental idea of using a kernel to choose $\phi(X_i)$ is that it will allow us to instead work with an N-dimensional vector K_i corresponding to each oservations, taken from the rows of a kernel matrix. Rather than having to construct $\phi(X_i)$ and seek equal means for the treated and control on it, as shown below it is sufficient to instead obtain equal means on K_i. This application of the "kernel trick" brings the power of machine learning methods that operate in highly non-linear, non-additive spaces to the problem of finding suitable balance. Below, I explain in greater detail why this substitution is possible.

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. I discuss the choice of b further below. It is common to rescale each covariate prior to computing $k(X_i, X_j)$, dividing by the standard deviation. This ensures results will be invariant to unit-of-measure decisions.

Two additional pieces of notation will be called upon. 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, the i^{th} row (or column) of $\mathbf{K}$ will be written as $K_i = [k(X_i, 1), k(X_i, 2), ..., k(X_i, N)]$. We will typically use K_i as an N dimensional vector describing observation i, thereby producing a richer representation of X_i.

2.5.2 Kernel as inner-product

For any kernel k 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 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 $k_{i,j} = \phi(X_i)^\top \phi(X_j)$. The generalization of this to potentially infinite-dimensional eigenfunctions is given by Mercer's Theorem (Mercer, 1909).

Note that 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. I describe the function space linear in these features in Section 5.3.

2.6 Mean Balance on K

This section defines mean balance in terms of $\mathbf{K}$ and introduces useful notation. To reduce 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 $\mathbf{K}$ for the treated can then be written $\frac{1}{N_t} \mathbf{K_t} \mathbf{1}_{N_t}$, while the weighted average row of $\mathbf{K}$ is $\mathbf{K_c} w$ for the $N_0 \times 1$ vector of weights w, with weights summing to 1.

Similar to mean balancing on X_i, 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:

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

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

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

2.7 Replacing $\phi(X_i)$ with K_i

The equality of $k(X_i, X_j)$ and $\langle \phi(X_i), \phi(X_j) \rangle$ may not at first seem to be a useful relationship, however this is precisely what makes it possible to show that balance on K_i is equivalent to balance on $\phi(X_i)$, despite K_i being only N-dimensional while the $\phi(X_i)$ associated with the Gaussian kernel is infinite-dimensional. This section offers two different approaches to showing this equivalence.

2.7.1 Equivalent spans of $\phi(X_i)$ and K_i

While Proposition 1 below will more directly show that mean balance on K_i achieves mean balance on $\phi(X_i)$, I first give a more intuitive result, showing that the space of functions given by $\phi(X_i)^\top\theta$ is also representable as $K_i c$ (for K_i as a row vector and c a column vector of N real coefficients). That is, the two basis sets have the same span, and we get just as much from mean balance on K_i as on $\phi(X_i)$. This justifies simply replacing $\phi(X_i)$ with K_i to estimate the desired weights. It will also help us to understand the function space spanned by $\phi(X_i)$ because the bases K_i are easier to understand (See 5.3).

To see this, 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

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

That is, the predictions linear in $\phi(X_i)$ are also linear in K_i for $\lambda > 0$. Because Y_{0i} is linear in $\phi(X_i)$ – and in K_i – equal means for the treated and control on K_i leads to equal means on Y_{0i}.

2.7.2 Direct equivalence of mean balance on K_i and $\phi(X_i)$

While the "equivalent spans" view above is valuable, particularly when we seek to understand this function space (see Section 5.3), I remark that a more direct route is available: mean balance on K_i directly implies mean balance on $\phi(X_i)$ as a result of the fact that $k(X_i, X_j) = \langle\phi(X_i), \phi(X_j)\rangle$.

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

Proof is given in Appendix A.4. Having given the treated and control groups the same mean on each dimension of $\phi(X)$, any function linear in $\phi(X_i)$ will have the same mean for the treated and control as well, without need of first showing the equivalence of $\phi(X_i)^\top\theta$ and $K_i c$ as above.

Under either motivation, the weights used to achieve mean balance on $\mathbf{K}$ are then employed in the weighted difference in means estimator, $\widehat{DIM}_w$ described in Theorem 1. This remains $\widehat{DIM}_w = \frac{1}{N}\sum_{i:D_i=1} Y_i - \sum_{i:D_i=0} w_i Y_i$, where now w_i are determined such that $\sum_{i:D_i=0} k_i w_i = \frac{1}{N_1}\sum_{i:D_i=1} k_i$ and subject to $\sum_i w_i = 1$ and $w_i > 0$, $\forall i$.

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

What remains is to choose weights, w_i that obtain mean balance on $\mathbf{K}$, and the procedure for choosing a lower-dimensional approximation of $\mathbf{K}$ that preserved as nearly as possible its linear span.

First, we have great flexibliity in the choice of weights, and in particular, a measure of divergence from uniform weights we wish to keep at a minimum subject to achieving the balance constraints in (Definition 1). 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).

Second, since the columns of $\mathbf{K}$ may be highly correlated, it is preferable to work with a lower-dimension approximation. One natural choice might be the rank-r approximation, $\widetilde{\mathbf{K}}^{(r)}$, closest to $\mathbf{K}$ in the Frobenius norm, i.e. minimizing

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

However, recall that our aim in achieving mean balance on $\mathbf{K}$ is to ensure balance 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 principal components of $\mathbf{K}$, we need not actually work with the projected approximation $\widetilde{\mathbf{K}}^{(r)}$ – we can simply work directly with the principal components themselves. 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. What remains is the choice of r, which can be chosen to minimize the resulting imbalance in $\mathbf{K}$. In practice, it is easy to obtain balance on r of sufficient size to capture well over 99% of the original variance of $\mathbf{K}$. We choose r so as to minimize an L_1 measure of imbalance described below. See Appendix A.1.2 for details, and Appendix A.6.2 for an illustration of this approach's performance in achieving balance on a non-linear function of the covariates.

I now turn to establishing this link between balance on $\mathbf{K}$ and equality of kernel estimates of multivariate density for the treated and controls.

2.9 Smoothed multivariate balance

The principle motivation for kernel balancing is as a reliable and hands-off method for estimation of the ATT (or ATC or ATE, see Section 5.5) 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, how does the procedure relate to methods such as matching that seek to make the multivariate density of the controls approximately equal to that of the treated? The use of kernels for the choice of $\phi(X_i)$ above produces a very useful equivalence: kernel balancing using kernel $k(\cdot,\cdot)$ implies that for a kernel density estimator also using kernel k, the multivariate density of the covariates so estimated is equal for the treated and control groups at all locations in the dataset. It thus also achieves (up to estimation limits due to dimension reduction on K_i) in a finite sample the goal of "multivariate balance" normally targeted by matching and weighting procedures, but only insofar as those densities are estimated using the same 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 anyhow, making this is a very useful equivalence. As a corrolary, 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 2 (Balance in $\mathbf{K}$ implies equality of smoothed multivariate densities) *Consider a density 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 2 is given in the appendix. Here I briefly build an intuition for this result, as it leads to further insights. First, the typical Parzen-Rosenblatt window approach estimates a density function according to:

$$\hat{p}(x) = \frac{1}{N\sqrt{4\pi b}} \sum_{i=1}^{N} k(x, X_i) \tag{8}$$

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 8 utilizing a Gaussian kernel generalizes to a multivariate density estimator based on Euclidean distances. Such density estimators are intuitively understandable as a process of placing a multivariate Gaussian kernel over each observation's location in $\mathbb{R}^P$, then summing them into a single surface and rescaling, providing a density estimate at each location.

The link between obtaining mean balance on Y_{0i} and obtaining multivariate density balancing emerges from the fact that both are manipulations of the superpositions of kernels placed over each observation. For a sample consisting of $X_1, \ldots, X_N$, construction of the kernel matrix $\mathbf{K}$ using the Gaussian kernel and right-multiplying it by a column vector, $\frac{1}{N\sqrt{4\pi b}}$, 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.

We would like to choose the weights such that the weighted, estimated density of the controls equals that estimated density of the treated, at every observation's location in covariate space. Using the formulas above, we can rewrite the estimated density o f 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 simply gives $\frac{1}{N_1}\sum_i K_i = \sum_i K_i w_i$ – which is simply mean balance on K_i again. That is, the moment conditions that produce mean balance on K_i are those that imply equal estimated distributions of the covariates when kernel k is used to make those estimates. This connection illuminates the deep connection between an assumption one makes on the outcome space of models and the sense in which multivariate density balance is achieved in ensuring balance on Y_{0i} in that space.

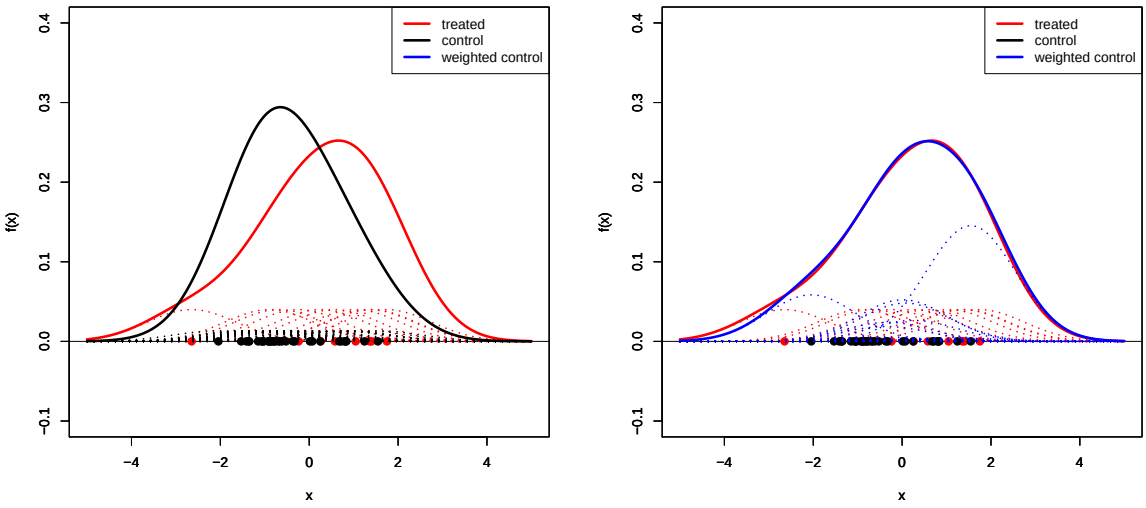
Finally, it is useful to introduce a measure of imbalance that can be used prior to or after weighting. Let us consider first our goal of achieving mean balance on K_i. To minimize imbalance in this sense, we might wish to minimize a p-norm proportional to $||\overline{k_t} - \sum_{i:D=0} w_i k_i||_p$. 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.1.3 for details. Here, I report the L_1 norm speficially, givne by $\frac{1}{2}||\frac{1}{N_1\sqrt{4\pi b}}\mathbf{K_t}^\top \mathbf{1}_{N_1} - \frac{1}{\sqrt{4\pi b}}\mathbf{K_c}^\top w||_1$, because it is naturally interpertable as an average of the gap between the 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 exacted matching, but here does not require coarsening the covariates into discrete bins as proposed there. Note that this is the same norm used choose dimensionality r for kernel matrix approximation $\mathbf{K}^{(r)}$: rank r is increased until the minmium of this norm is found (Appendix A.1).

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} that is essential for unbiasedness, and which kernel balancing targets.

3 An Illustration: Effect of peacekeeping and imbalance on a ratio

The simulation shown here 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

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

Sambanis, 2000). However, within the set of civil war cases constituting our sample, the "treatment" – peacekeeping missions (*peacekeeping*) – is not ra ndomly 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 backed out, using $intensity \cdot war\ duration$.

Suppose the outcome of interest, *peace years*, is also a function of *intensity*, with more intense

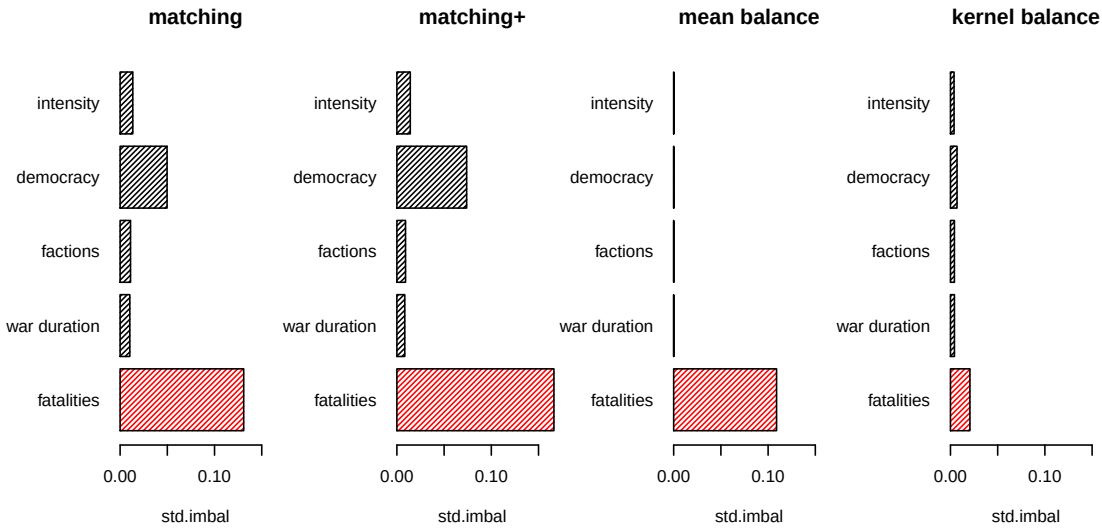
conflicts leading to longer average *peace years* according to

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

where the term $-(0.5)peacekeeping_i$ 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.

Figure 2: Simulation: balance on an important but unknown function of covariates

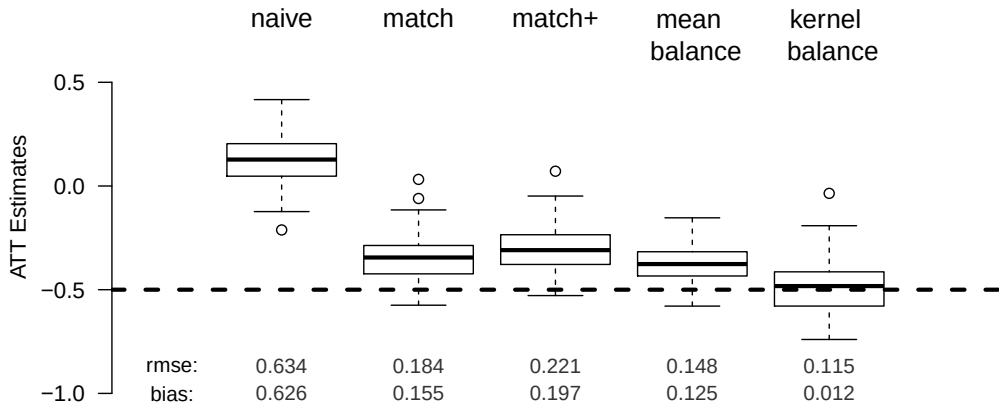


Mean imbalances on included covariates and $intensity = \frac{fatalities}{war\,duration}$, which determines both assignment of the treatment (*peacekeeping*) and the outcome (*peaceyears*). *Matching:* Mahalanobis distance matching on the original covariates alone leaves a substantial imbalance on *war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

How well do existing techniques achieve equal means for the treated and controls ("mean balance"), both on the original four covariates and on *intensity*, a (non-linear) function of the observables? In Figure 2, the horizontal axis for each plot shows the standardized difference in means between treated and control on each of the covariates, as well as on *intensity*. All results are taken over 500 simulations with the same data generating process and $N = 500$ on each. 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 provide the needed flexibility to produce balance on *intensity*. In fact, balance on both *war duration* and *intensity* are worsened. In the third plot (*mean balance*), entropy balancing (Hainmueller, 2012) is used to achieve equal means in the original covariates. As expected, this produces excellent balance on the original covariates, but only a modest improvement in balance on *intensity*. Finally, the fourth 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*.

Figure 3: Simulation: distribution of effect estimates by method



Boxplot illustrating distribution of average treatment effect on the treated (ATT) estimates in the same example as Figure 2 above. The actual effect is -0.5 *peace years*. *Matching*, *matching+*, and *mean balance* all show large biases because the control samples chosen by these procedures include higher *intensity* conflicts than the treated sample, even though *intensity* is entirely a function of observables. Since *intensity* influences the outcome, *peace years*, the treated and control samples thus differ regardless of any treatment effect. By contrast, *kernel balance* is approximately unbiased, as it achieves balance on a large space of smooth functions of the covariates.

These imbalances are worrying because they indicate a failure to condition on the covariates as required to achieve ignorability, and lead to biased ATT estimates. When the ATT is estimated by difference in means in the post weighting/matching sample, large bias occurs with the exception of kernel balancing (Figure 3). Kernel balancing shows the lowest bias by far among the methods attempted. It's advantages in RMSE are mode modest, but it still has 22% lower RMSE than the next best estimator, *mean balance*.

Finally, while kernel balancing is largely automated and avoids the need for iterative specification or balance testing, given a choice of Gaussian kernel one still chooses the bandwidth parameter, b. Section 5.4 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 5.4 for discussion of this default value).

This illustration while simple and artificial, demonstrates the ease with which existing methods can

fail: a non-linear function of two observed covariates – even a simple ratio – may influence the potential outcomes, producing scenarios in which existing matching and weighting methods pose risks of large biases. An investigator's theoretical knowledge is rarely if ever 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.

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 a "true" answer is known. 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 (e.g. Diamond and Sekhon, 2005; Iacus et al., 2011; Hainmueller, 2012).

The aim of these studies is to recover an experimental estimate of the effect of a job training program, the National Supported Work (NSW) program. Following LaLonde (1986), the treated sample from the experimental study is compared to a control sample drawn from a separate, observational sample. Methods of adjustment are tested to see if they accurately recover the treatment effect despite large observable differences between the control sample and the treated sample. See (Diamond and Sekhon, 2005) for an extensive description of this dataset and the various subsets that have been drawn from it. Here I use 185 treated units from NSW, originally selected by Dehejia and Wahba (1999) for the treated sample. The experimental benchmark for this group of treated units is \$1794, which is computed by difference-in-means in the original experimental data with these 185 treated units. The control sample is drawn from the Panel Study of Income Dynamics (PSID-1), containing 2490 individuals.

The pre-treatment covariates available are age, years of education, real earnings in 1974, real earnings in 1975 and a series of indicator variables: Black, Hispanic, and married. However, as this dataset has now been used many times, it is common practice to use three further variables that are actually non-linear transforms of these: indicators for being unemployed (having income of \$0) in 1974 and 1975, and an indicator for having no highschool degree (fewer than 12 years of education).

As found by Dehejia and Wahba (1999), propensity score matching can be effective in recovering reasonable estimates of the ATT, but these results are highly sensitive to specification choices in constructing the propensity score model (Smith and Todd, 2001). Diamond and Sekhon (2005) use genetic matching to estimate treatment effects with the same treated sample. While matching solutions with the highest degree of balance produced estimates very close to the experimental benchmark, these models included the addition of squared terms and two-way interactions, 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.

Figure 4 reports results from a variety of estimation procedures and specifications. Three procedures are used: linear regression (*OLS*), Mahalanobis distance matching (*match*), and kernel balancing (*kbal*). For *match* and *kbal*, estimate are produced by simple difference in means on the matched/reweighted sample. For comparability, all three approaches use simple standard errors that ignores any pre-processing stage, i.e. the usual "fixed" weight standard errors.

For each method, three sets of covariates are attempted: the "standard" set of 10 covariates described above; a reduced set (*simple*) including only the seven of these that are "original" variables, not transforms of others; and an expanded set (*squares*) including the 10 standard covariates plus squares of the three continuous variables.

Figure 4 shows that the OLS estimates vary widely by specification, and even the estimate closest to the benchmark of $1794 is incorrect by $1042. Mahalahobis distance matching performs better, though remains somewhat specification dependent, with its best estimate (*match-squares*) falling within $387 of the benchmark. Finally, kernel balancing (*kbal*) performs well over the three specification, with no estimate more than $681 from the benchmark, and specification using the standard covariate set producing an estimate of $1807, $13 from the benchmark.

Perhaps the most important benefit of kernel balancing in this example is its relative insensitivity to specification. Matching performs well when the researcher knows to seek balance on particular nonlinear functions – specifically, on "unemployment" in 1974 and 1975, which are actually indicators for zero income. This transform of the original covariates is included in both *match_squares* and *match* specifications. However in the *match_simple*, where only untransformed variables are used, matching produces a very different estimate, with a confidence interval that actually excludes the benchmark. By contrast, 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 (where all methods use the "fixed weights" approach for comparability here) 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).

Further investigating the kernel balancing solution reveals additional details. We can see that balance is difficult to achieve in this example, in the sense that it requires focusing on a relatively small portion of the original control sample. At the solution achieved by kernel balancing on the original variables alone (*kbal_simple*), 90% of the weight for controls is taken from just 98 units (reported automatically by `kbal`). So much weights falls to so few observations due to large differences between the treated and control samples. In examining why this is, the unemployment variable reveals its value: while 72% of the treated are unemployed in either 1974 or 1975, only 12% of controls are unemployed in either year. Using the L_1 measure described above, interpetable as either remaining imbalance on $\mathbf{K}$ or smoothed multivariate density imbalance, we get a value of $L_1 = 0.41$ prior to weighting. This indicates a considerable gap between the heights of the (smoothed) densities of the treated and control as evaluated at each datapoint. This is reduced to just $L_1 = 0.0016$ by kernel balancing. The choice of r selected by the procedure was 45 dimensions. With this choice, principal components accounting for 99.7% of the total variance of $\mathbf{K}$ are balanced upon.

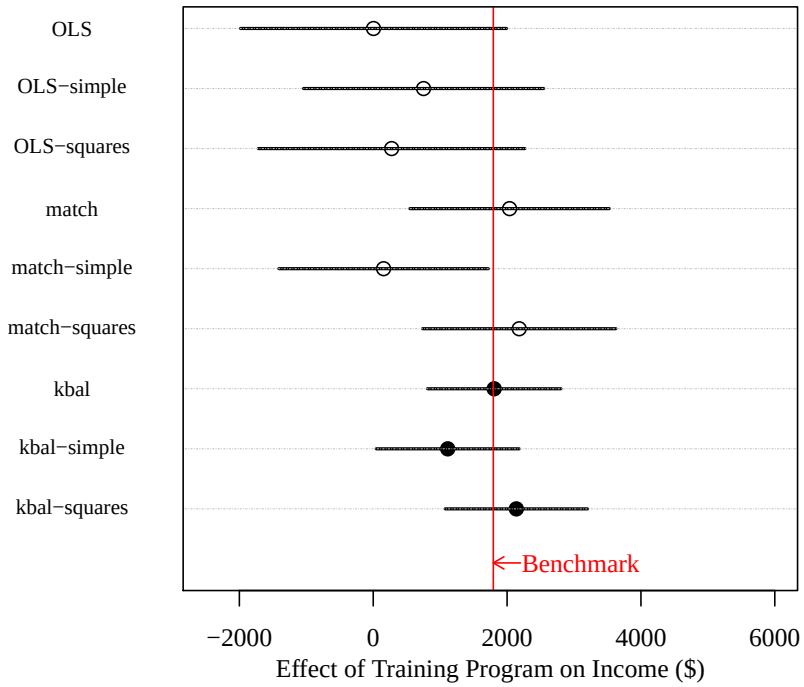
5 Discussion

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

5.1 Relation to Existing Approaches

Here, I compare kernel balancing to matching, covariate balancing weights, and propensity score methods. Like kernel balancing, each of these begins with an ignorability assumption (Assumption 1).

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



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

However, these methods exploit Assumption 1 to make causal inferences through the more difficult estimation route of seeking multivariate balance rather than merely balance on Y_0. 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, treatment assignment is independent of potential outcomes within each stratum of X. The most natural way to exploit this for estimating the SATT is to perform this conditioning on X very literally: take difference-in-means estimates of the treatment effect within each stratum of X, then average these together over the empirical distribution of X for the treated. Subclassification and exact matching estimators do this. However, conditioning on X in this way is 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 is incomplete, and the distribution of X for the treated and controls are not 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 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. In short, the goal of matching is to align the multivariate distribution of covariates for the control units with that of the treated, but matching discrepancies can prevent this from occurring, and the tools used to test for this multivariate balance are incomplete. As the motivating example in Section 3 illustrates, matching can thus fail to obtain sufficient similarity of distributions, even when investigators attempt to match on higher-order terms.

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), with a number of similar procedures emerging from the survey sampling literature, such as raking (Kalton, 1983). Once these moment conditions are satisfied, it is assumed that the multivariate densities for the treated and control are alike in all important respects. These weights can be used in a difference in means estimation or other procedure. The upside of this procedure over matching is that the prescribed moments of the control distribution can often be made exactly equal to those of the treated, avoiding the matching discrepancy problem. The downside is that it loses the non-parametric quality of matching, providing balance only on enumerated moments. It is generally not possible to know what moments of the distribution must be balanced to ensure unbiasedness, because we do not know which functions of the covariates might influence the (non-treatment) outcome. Kernel balancing can be understood as an extension to these covariate balancing weighting methods that solves this problem by ensuring balance on a large space of smooth functions of the covariates automatically.

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.7 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. As also shown there, these weights can be rewritten via Bayes rule as the ratio of class densities for the treated and controls,

$$w_{IPW} = \frac{p(x|D_i = 1)}{p(x|D_i = 0)} \tag{9}$$

Written in this way, it becomes clear that whenever the class densities are equal for the two groups, the IPW weights would have to remain constant at 1. This makes sense, since two classes with identical multivariate distributions would indeed be indistinguishable, producing constant propensity scores under a generative model for the probability of taking the treatment. Given the multivariate balancing property discussed above, kernel balancing weights achieve precisely this equality of class densities, insofar as multivariate density is estimated by the corresponding kernel density estimator (Section 2.9). This provides an intimate relationship between kernel balancing and inverse propensity score weighting: inverse propensity score weights become constant (and thus unnecessary) in a sample that has been weighted by kernel balancing already, but only when the corresponding kernel density estimator is used. Yet, kernel balancing does not explicitly model a propensity score.

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.

Kernel balancing assumes the existence of, but does not estimate, an outcome model $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top \theta = K_i^\top \mathbf{c}$. Rather than fitting such a model or otherwise utilizing the outcome data, kernel balancing uses this assumed existence as a device for determining what basis functions need to have the same mean for the treated and control groups in order to ensure that the $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 0]$. Indeed, once two samples have equal means on $\phi(X)$ (or equivalently, K_i), no outcome model need be employed to estimate the ATT – difference in means is sufficient. That said, double-robust estimation options that further include covariates and an outcome model on the weighted data can be considered as well (see Zhao and Percival, 2016 for double-robustness and optimal efficiency results in the special case where entropy balancing is used to achieve mean balance directly on X and Y_{0i} is assumed linear in X).

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 the value of θ or $\mathbf{c}$, and we do not need to rely on the accuracy of estimates for these quantities 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 and more importantly, performing a weighting approach whose justification is rooted in a choice of outcome models is not equivalent to using the outcome model alone, because the former changes the distribution of (in this case) the control group to be more similar to that of the treated prior to estimation of an effect. Such a "pre-processing" approach (Ho et al., 2007) is very helpful: once the treated and control groups are made similar in their characteristics through this reweighting, the investigator no longer requires heroic modeling assumptions to bridge the gap between treated and control units that may lie far apart in the covariate space. This point is not a trivial one, and the results in the empirical example above (Section 4) provide one dramatic illustration. There, treated and

control distributions differ radically in their distribution of the covariates and simply using an outcome model with covariates and a treatment indicator does a very poor job of estimating how differences in the covariates should be used to implicitly adjust the outcomes over such a wide range of X. As a result it estimates small or even negative effects of a job training program on income, while matching and weighting estimators reveal a positive effect that closely matches the experimental result.

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, a common 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 esimated 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 do 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 is likely to be valid. While further work is needed on more computationally attractive alternatives, boostrapping the entire procedure of selecting weights by kernel balancing then estimating the subsequent treatment effect is likely an appropriate choice for users who wish to incorporate uncertainty from the weight selection into the final estimates.

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

One reason to use the Gaussian kernel is that it is widely used as a workhorse kernel in machine learning regression and classification tasks, owing to the flexibility and generality of the correpsonding function space. Though we are not actually fitting a model here, the function space invoked, $\phi(X_i)^\top\theta$, is the same. Relatedly, for the Gaussian kernel this feature space has the universal representation property: as $N \to \infty$, $\phi(X)^\top\theta$ can fit any continuous function of X (Micchelli et al., 2006).

Of course, this universality as N approaches ∞ is less reassuring in small 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. Further justification for the Gaussian choice of kernel and an intuition for the nature of this function space is given in section 5.3. While it is unconventional to describe the function space associated with the Gaussian kernel in such detail, I do so here as investigators must be able to ascertain whether the assumption that $\mathbb{E}[Y_{0i}|X_i]$ lies in this space is reasonable in their application.

A key assumption of the method is that $\mathbb{E}[Y_{0i}|X_i]$ can be well fitted by $\phi(X_i)\theta$ (Assumption 2), where $\phi(\cdot)$ is determined by a particular choice of kernel. In this implementation, I focus on the Gaussian kernel, and so it is useful to understand what this function space looks for this choice.

This function space is the same Reproducing Kernel Hilbert space of functions used by numerous regression and classification methods employing a Gaussian kernel, including kernel ridge regression, support vector machines, and Gaussian processes. 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. In fact the choice of $\phi(X)$ such that $\langle \phi(X_i), \phi(X_j) \rangle = k(X_i, X_j)$ is not unique. One valid choice for $\phi(X)$ in the case of the Gaussian kernel is the sequence given by $\left\{ \sqrt{\frac{2^d}{d!}} exp(-X_i^2)(X_i)^d \right\}$ for $d = 0, 1, ..., \infty$ (see Appendix A.6).

More usefully, however, as shown above the functions linear in $\phi(X_i)$ are also those linear in K_i (note that $k(X_i, \cdot)$ is sometimes called the "canonical feature mapping" corresponding to $\phi(x)$). 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 kernels 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 highly non-linear but smooth functions. More generally, this machinery is at work in any kernel machine using a Gaussian kernel, including support vector machines, radial basis function neural networks, and Gaussian processes.

This space of functions is appealing because while making no assumptions of linearity or additivity in X, it is generally reasonable to assume that the conditional expectation of Y_{0i} is continuous and relatively smooth over $\mathcal{X}$. As noted above, this feature space has universal representation property, such that as $N \to \infty$, $\phi^\top(X)\theta$ can fit any continuous function of X (Micchelli et al., 2006). While less reassuring in small samples, the superposition of Gaussians views makes clear that smoother functions can be fitted with fewer observations, making this an excellent choice to model $\mathbb{E}[Y_{0i}|X_i]$ when little is known about the nature of the relationship except that it is continuous and likely to be smooth. Accordingly, the Gaussian kernel is the "workhorse" choice for many kernelized regression and classification models.

By achieving equal expectations for the treated and control on the columns of $\mathbf{K}$, kernel balancing thus ensures that the many smooth functions that can be built by the superposition of Gaussians will have the same mean for the treated and control group. It is thus suitable when we have little knowledge of the shape of $\mathbb{E}[Y_{0i}|X_i]$ but believe it is well approximated in such a flexible functions space.

5.4 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. Here we consider the scaling and rotation of X, and the choice of b.

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 largest 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 it is much more important to choose b according to how it effects mean balance in Y_{0i}. To this end, the choice of parameter b is a feature-extraction decision that determines the construction of $\phi(X_i)$ and thus $\mathbf{K}$. It determines how close two points X_i and X_j need to be in order to have highly similar rows k_i and k_j. This implies a bias-variance tradeoff. If b is too large, mean balance is easier to achieve and the weights will have low variance, the resulting balance is less precise (and the corresponding smoothed densities more "blurred"). If b is too small, $\mathbf{K}$ will approximate the identity matrix, and each row k_i will be nearly linearly independent. In this case, the algorithm will not converge as balance cannot be attained. (The possibility of trimming away treated units that are difficult to match under small b is discussed in Appendix A.8).

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

One reasonable choice that may be a useful reporting standard would be to use $b = dim(X)$, while showing results at other choices for robustness. 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 very good performance. This is the default value of b used here, though clearly further work is needed on this point. Investigators may wish to present their results across a range of b values to ensure this choice is not consequential in a given application. Shoudl the results vary across b values, inspecting L_1 and the concentration of weights (e.g. through $min90$) can be helpful for determining an appropriate value.

5.5 Other Quantities: ATE, ATC

I have focused on the ATT for simplicity of exposition, but with minor adjustment this method can also be used to identify the average treatment effect on the controls (ATC) and the average treatment

effect on the treated.

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. Note that identification of the ATC requires an assumption analogous to 1 but for both non-treatment outcomes, specifically $Y_{1i} \perp\!\!\!\perp D_i | X_i$. The ATE requires ignorability with respect to both of the potential outcomes conditionally on X_i, $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i | X_i$.

6 Conclusions

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

However, unbiasedly estimating the SATT requires only that $\frac{1}{N_1} \sum_{i:D_i=1} Y_{0i} = \frac{1}{N_1} \sum_{i:D_i=0} Y_{0i}$, or "mean balance on Y_{0i}". Kernel balancing achieves this goal by working with the kernel matrix, $\mathbf{K}$, rather than the original covariates, X. It finds weights on the controls to make the weighted average row of $\mathbf{K}$ for the controls equal to the average row of $\mathbf{K}$ for the treated. Mean balance on these features implies mean balance on a large set of smooth functions of X. This includes all functions that can be formed by the superposition of Gaussians placed over each observation in the covariate space – a very flexible space of functions that fits smooth functions particularly well in even smaller samples. The assumption that $\mathbb{E}[Y_{0i}|X_i]$ is among these functions is far more plausible than the assumption that it is linear in the original X, even if the investigator is careful enough to include higher-order terms among these X's. Moreover as N grows large, $\mathbb{E}[Y_{0i}|X_i]$ can be increasingly well accomodaed in this space.

While mean balance on Y_{0i} is the principle goal, kernel balancing also implies that a particular

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

Kernel balancing performs well in a reanalysis of Dehejia and Wahba (1999), a widely used benchmark for covariate adjustment in causal inference. At it's default values, with the covariates commonly used for this problem and no further specification choices, kernel balancing estimated an effect of $1807 using the non-experimental control group, extremely close to the experimental benchmark of $1794. Moreover, results are stable across specifications: getting an accurate result dow not depend upon foreknoweldge of non-linear functions that must be included to get a good result. Here agin, kernel balancing can be thought of as a principled choice of what functions of the covariates to acheive mean balance on, such that resulting ATT estimates are unbiased even when we do not know how exactly the covariates may influence the (non-treatment potential) outcome.

Numerous questions and challenges 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 is likely valid for obtaining confidence intervals that include uncertainty due to weight selection, further work on this is needed, and particularly on any 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, alternate methods for dimension reduction on $\mathbf{K}$, and alternative methods for choosing the balancing weights that achieve mean balance on $\mathbf{K}$ while minimizing volatility. An implemnetation of this procedure using the choices described here is available in the `R` package `KBAL`, to be distributed on the CRAN repository upon acceptance of this paper.

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.

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.

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.

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., 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.

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

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.

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

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

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.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). 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.

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.

Zhao, Q. and Percival, D. (2016). Entropy balancing is doubly robust. *Working Paper*.

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

A 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 Optimization over r

As described in the text, balance is achieved on the first r principal components of $\mathbf{K}$, thereby achieving balance on the reconstructed approximation $\widetilde{\mathbf{K}}^{(r)}$ closest to $\mathbf{K}$ in the Frobenius norm and operator 2-norm senses. How should r be chosen? Since the goal is to achieve $\overline{k_t} = \sum_{i:D=0} w_i k_i$, a natural imabalnce measure to judge the success of a set of weights would be $a||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ for some norm $||\cdot||$ and constant a. Following (Iacus et al., 2011), I choose the L_1 measure here. As discussed below, this can serve as a measure of both imbalance on $\mathbf{K}$ and multivariate density imbalance insofar as densities are estimated by the corresponding kernel smooth (see A.1.3).

The choice of r is then made by beginning with $r = 1$ and increasing it until a minimum in imbalance as measured by $L_1 = \frac{1}{2}\sum_{i=1}^{N} |\overline{k_t} - \sum_{i:D=0} w_i k_i|^1$. Alternative choices of norm (such as L_2 produce very similar results). Typically, imbalance improves as r initially rises, and then deteriorates once r is too high and numerical instability begins to creep in. An illustration of the relationship r, L_1 and the balance achieved on unknown functions of X is given in the appendix (Figure 7). In practice, for the r chosen in this way, the number of principal components balanced upon generally accounts for well over 99% of the variance of $\mathbf{K}$.

A.1.3 Equivalence of K-imbalance and smoothed multivariate density imbalance

Recall that the choice he 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.

$$SATT = \frac{1}{N_1} \sum_{i:D_i=1} Y_{1i} - \frac{1}{N_0} \sum_{i:D_i=0} Y_{0i} \tag{10}$$

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

$$\widehat{DIM}_w - SATT = \frac{1}{N_1} \sum_{i:D_i=1} Y_{0i} - \sum_{D_i=0} w_i Y_{0i} \tag{11}$$

$$= \frac{1}{N_1} \sum_{i:D_i=1} \left(\phi(X_i)^\top \theta + \epsilon_i \right) - \sum_{i:D_i=0} w_i \left(\phi(X_i)^\top \theta + \epsilon_i \right) \tag{12}$$

$$= \theta^\top \frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \theta^\top \sum_{i:D_i=0} w_i \phi(X_i) - \sum_{i:D_i=0} w_i \epsilon_i \tag{13}$$

$$= \theta^\top \left(\frac{1}{N_1} \sum_{i:D_i=1} \phi(X_i) - \sum_{i:D_i=0} w_i \phi(X_i) \right) + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i \tag{14}$$

$$= 0 + \frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i \tag{15}$$

The bias is the expectation of this quantity,

$$bias = \mathbb{E}\left[\widehat{DIM}_w - SATT \right] \tag{16}$$

$$= \mathbb{E}\left[\frac{1}{N_1} \sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i \epsilon_i \right] = 0 \tag{17}$$

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 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

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i + \frac{1}{N_1}\sum_{i:D_i=1} h(X_i) - \sum_{i:D_i=0} w_i h(X_i) \tag{18}$$

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 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:

$$\begin{aligned}
\mathbb{E}[Y_{0i}|D=1] &= \mathbb{E}_x\left[\mathbb{E}[Y_{0i}|X, D=1]\right] && (19)\\
&= \theta^\top \int \phi(x)p(x|D=1)dx && (20)\\
&= \theta^\top \mathbb{E}[\phi(x)|D=1] && (21)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_w[Y_{0i}|D=0] &= \mathbb{E}_{w,x}\left[\mathbb{E}[Y_{0i}|X, D=0]\right] && (22)\\
&= \theta^\top \int \phi(x)wp(x|D=0)dx && (23)\\
&= \theta^\top \mathbb{E}_w[\phi(x)|D=0] && (24)
\end{aligned}$$

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.4 Proof of proposition 1

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

This can be shown as follows.

$$\overline{k_T} = \sum_{i:D_i=0} w_i k_i \tag{25}$$

$$\frac{1}{N_1}\left[\sum_{i:D_i=1} k(X_i, X_1), \ldots, \sum_{i:D_i=1} k(X_i, X_N)\right] = \left[\sum_{i:D_i=0} w_i k(X_i, X_1), \ldots, \sum_{i:D_i=0} w_i k(X_i, X_N)\right] \tag{26}$$

$$\frac{1}{N_1}\sum_{i:D_i=1}[\langle\phi(X_i), \phi(X_1)\rangle, \cdots, \langle\phi(X_i), \phi(X_N)\rangle] = \sum_{i:D_i=0} w_i[\langle\phi(X_i), \phi(X_1)\rangle, \cdots, \langle\phi(X_i), \phi(X_N)\rangle] \tag{27}$$

$$\frac{1}{N_1}\sum_{i:D_i=1}\langle\phi(X_i), \phi(X_j)\rangle = \sum_{i:D_i=0} w_i\langle\phi(X_i), \phi(X_j)\rangle, \ \forall j \tag{28}$$

$$\langle\frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i), \phi(X_j)\rangle = \langle\sum_{i:D_i=0} w_i\phi(X_i), \phi(X_j)\rangle \tag{29}$$

$$\langle\overline{\phi_t}, \phi(X_j)\rangle = \langle\sum_{i:D_i=0} w_i\phi(X_i), \phi(X_j)\rangle \tag{30}$$

$$\overline{\phi_t} = \sum_{i:D_i=0} w_i\phi(X_i) \tag{31}$$

A.4.1 Remarks

An intuitive interpretation of equation 29 is that each unit j is as close to the average treated unit as it is to the (weighted) average control unit, where distance is measured in the feature space $\phi(X)$. For the Gaussian kernel, $\langle\phi(X_i), \phi(X_j)\rangle$ is naturally interpretable as a similarity measure in the *input* space, since this quantity equals $k(X_j, X_i) = e^{-\frac{||X_j - X_i||^2}{b}}$. However, $\langle\phi(X_i), \phi(X_j)\rangle$ or $k(X_i, X_j)$ is more generally interpretable as similarity in the feature space as well. Note the squared Euclidean distance between two points X_i and X_j after mapping into $\phi(\cdot)$ is: $\|\phi(X_i) - \phi(X_j)\|^2 = \langle\phi(X_i) - \phi(X_j), \phi(X_i) - \phi(X_j)\rangle = \langle\phi(X_i), \phi(X_i)\rangle + \langle\phi(X_j), \phi(X_j)\rangle - 2\langle\phi(X_i), \phi(X_j)\rangle$. In the case of the Gaussian kernel, $\langle\phi(X_i), \phi(X_i)\rangle = 1$, so this distance reduces to $2(1 - \langle\phi(X_i), \phi(X_j)\rangle)$. In this sense, $\langle\phi(X_i), \phi(X_j)\rangle$ is as reasonable measure of similarity of position in the feature space, as it runs opposite to distance in this space.

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

A.5 Proof of proposition 2

Proposition 2 states that for a density estimator for the treated, $\hat{f}_{X|D=1}$, and for the (weighted) controls, $\hat{f}_{X|D=0,w}$, both constructed with kernel k with scale 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:

$$\begin{aligned}
\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \frac{1}{\sqrt{2\pi b}}\mathbf{K_c}w \\
\frac{1}{N_1}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \mathbf{K_c}w \\
\overline{k_t} &= \mathbf{K_c}w \\
\overline{k_t} &= \overline{k_c(w)}
\end{aligned} \tag{32}$$

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.6 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:

$$k(X_i, X_j) = exp(-||X_i - X_j||^2/1) \tag{33}$$

$$= exp(-X_i^2)exp(-X_j^2)exp(2X_i, X_j) \tag{34}$$

$$= exp(-X_i^2)exp(-X_j^2)\sum_{d=0}^{\infty}\frac{2^d X_i^d X_j^d}{d!} \tag{35}$$

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

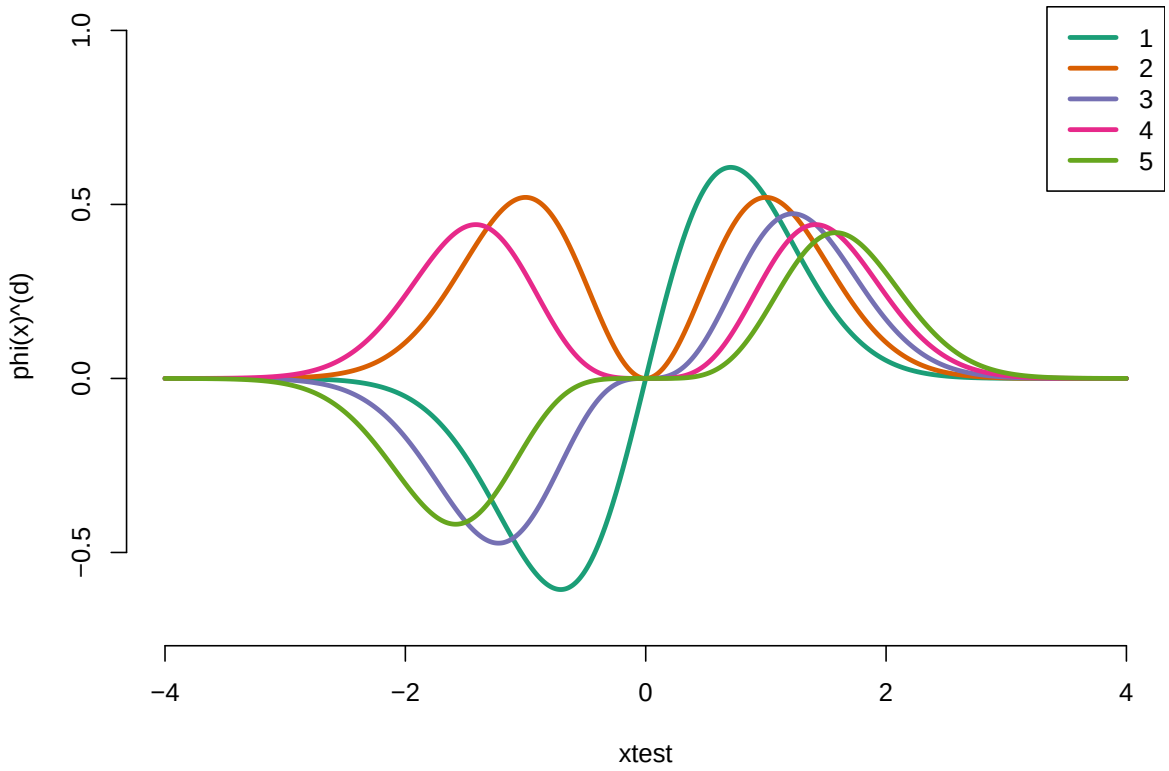
$$k(X_i, X_j) = \sum_{d=0}^{\infty}\sqrt{\frac{2^d}{d!}}exp(-X_i^2X_i^d)\sqrt{\frac{2^d}{d!}}exp(-X_i^2X_i^d) \tag{36}$$

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

$$\phi(X_i) = \left[\sqrt{\frac{2^0}{0!}}exp(-X_i^2X_i^0),\ \sqrt{\frac{2^1}{1!}}exp(-X_i^2X_i^1),\ \ ...,\ \ \sqrt{\frac{2^\infty}{\infty!}}exp(-X_i^2X_i^\infty)\right] \tag{37}$$

Figure 5 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.

Figure 5: 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.6.

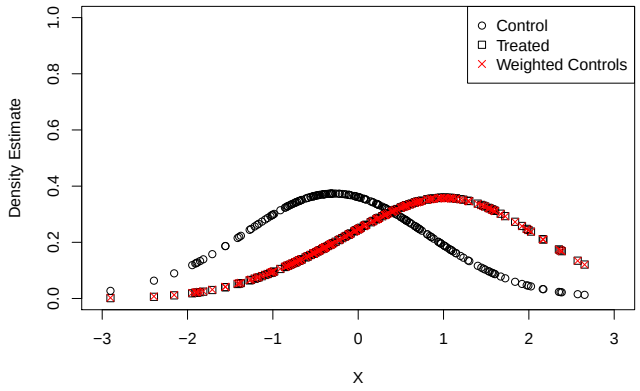
A.6.1 Density Equalization Illustration

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 6 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 $\frac{1}{N_0\sqrt{2\pi b}}\mathbf{K_c}\mathbf{1}_{N_0}$. As shown, the density estimates for the treated at each observations X position (black squares) is initially very different from the density estimates for the controls taken at each observation (black circles). After weighting, however, the new density of the controls as measured at each observation (red `x`) matches that of the treated almost exactly.

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

Figure 6: Density-Equalizing Property of Kernel Balancing



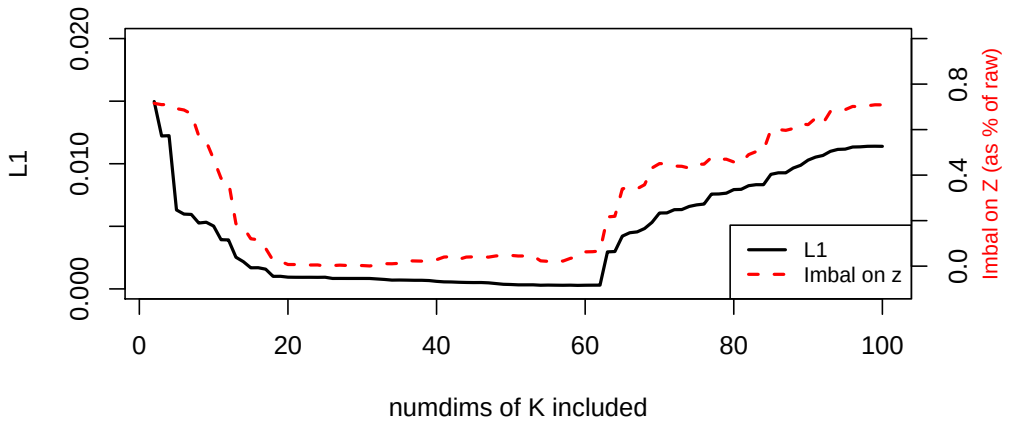
Plot showing the density-equalization property of kernel balancing. For 200 observations of $X \sim N(0,1)$, treatment is assigned according to $Pr(treatment) = 1/(1+exp(2-2X))$, producing approximately two control units for each treated unit. Black squares indicate the density of the treated, as evaluated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

A.6.2 L_1, imbalance, and r

Recall that kernel balancing does not directly achieve mean balance on $\mathbf{K}$, but rather on the first r factors of $\mathbf{K}$ as determines by principal components analysis. This example examines the efficacy of this approach in minimizing the L_1 loss, and in minimizing imbalance on an unknown function of the data. Suppose we have 500 observations and 5 covariates, each with a standard normal distribution. Let $z = \sqrt{x_1^2 + x_2^2}$. This function impacts treatment assignment, with the probability of treatment being given by $logit^{-1}(z-2)$, which produces approximately two control units for each treated unit.

In Figure 7, the value of r – the number of factors of $\mathbf{K}$ retained for purposes of balancing – is increased from a minimum of 2 up to 100. As expected, both L_1 and the mean imbalance on z taken after weighting improve as r is first increased, and then worsen beyond some choice of r. Most importantly, while the balance on z is unobservable in the case of unknown confounders, L_1 is observable, and improvements in L_1 track very closely to improvements in the balance of z. Accordingly, selecting r to minimize L_1 appears to be a viable strategy for selecting the value that also minimizes imbalance on unseen functions of the data.

Figure 7: L_1 distance and imbalance on an unknown confounder, by r



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

A.7 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:

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=1] && (38)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int \mathbb{E}[Y_{0i}|D_i=1,x]p(x|D_i=1)dx && (39)\\ &= \int \mathbb{E}[Y_{1i}|D_i=1,x]p(x|D_i=1)dx - \int \mathbb{E}[Y_{0i}|D_i=0,x]p(x|D_i=1)dx && (40) \end{aligned}$$

Expression 40 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,

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

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,

$$\text{DIM}_w = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}_w[Y_{0i}|D_i = 0] \tag{43}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int w_i \mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 0)dx \tag{44}$$

where w_i is a function of X that allows us to upweight or downweight control units. The difference between expression 40 and 42 can be resolved by choosing weights

$$w_i = \frac{p(x|D_i = 1)}{p(x|D_i = 0)} \tag{45}$$

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

Note that in the form 9, 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.8 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.

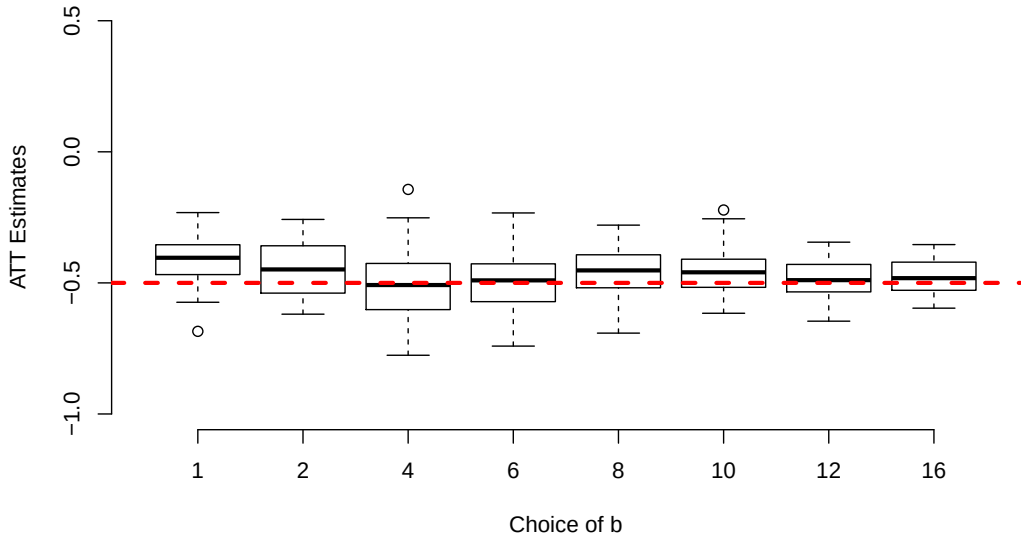
However, even when the investigator is willing to change the population of interest by trimming the treated, it is not always clear on what basis trimming should be done. In kernel balancing, trimming of the treated can be (optionally) employed by using the multivariate density interpretation given above. 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.9 Stability Across b in Simulation 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 $b = dim(X)$. Here we find 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.

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.

A.10 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.

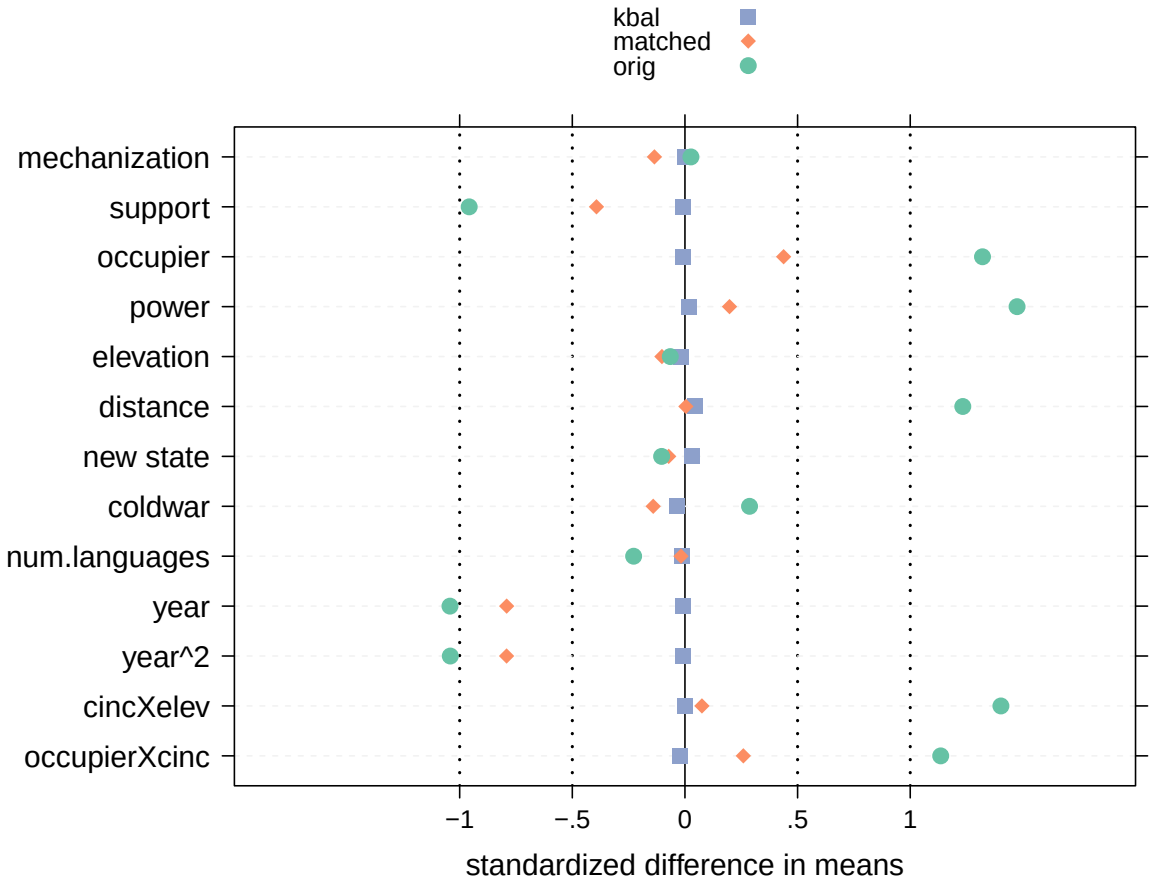
While the credibility of this estimate as a causal quantity depends on the absence of unobserved confounders, we can nevertheless assess whether the procedures used to adjust for observed covariates were sufficient, or whether an inability to achieve mean balance on some functions of the covariates may 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 9, 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

Figure 9: Balance: Democracies vs. Non-democracies and the Counterinsurgencies they Fight



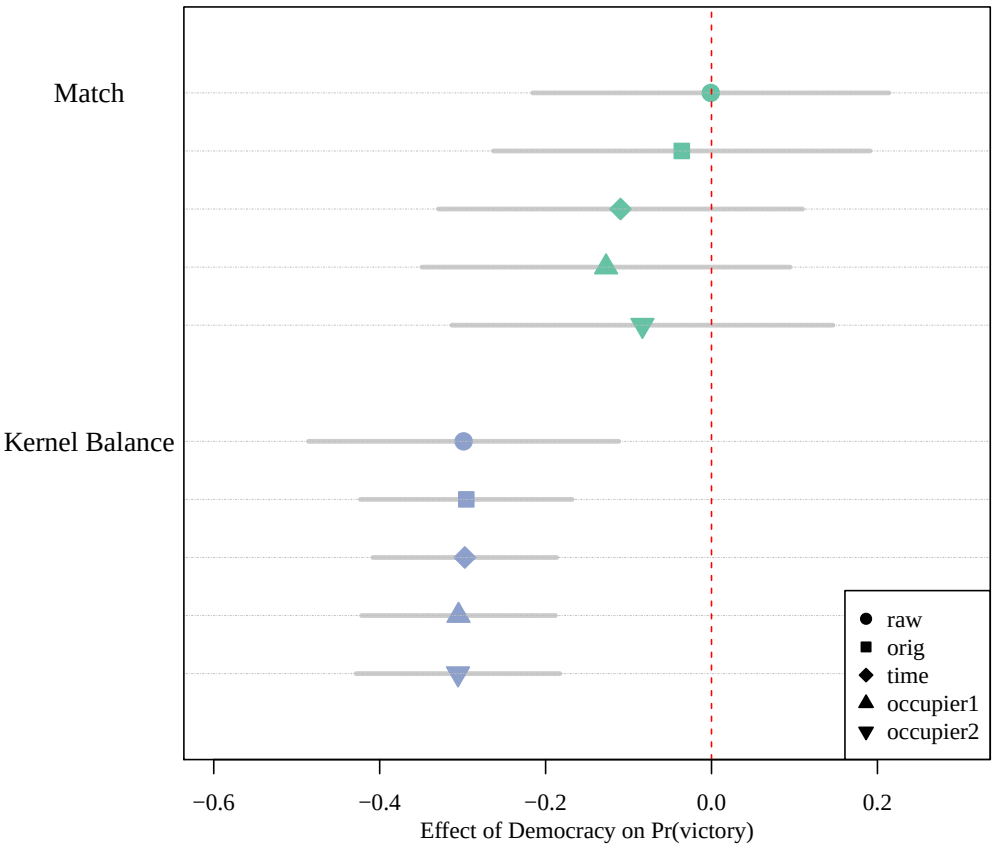
Balance in post-1945 sample of Lyall (2010). Imbalance, measured as the difference in means divided by the standard deviation, is shown on the x- axis. Democracies (treated) and non-democracies (controls) vary widely on numerous covariates. The matched sample (diamonds) shows somewhat improved balance 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.

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 10 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.

Figure 10: 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.

A.11 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.