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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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

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

May 29, 2019

Abstract

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

1 Introduction

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

By comparison, standard weighting approaches can (when feasible) achieve an exact or approximate balance on desired moments, such as the means of X_1 and X_2, and so may seem to avoid this problem. However, they do so by sacrificing the nonparametric quality of matching. If a weighting estimator obtains equal means for the treated and control groups on both X_1 and X_2 (referred to here as "mean balance"), this does not, in general, imply that X_1/X_2 or other nonlinear functions of X_1 and X_2 have equal means for the two groups, potentially resulting in bias. In short, both weighting and matching, as described thus far, would fail to make the treated and control groups "comparable" on the observables

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

for this simple example. The desired adjustment for the observables is thus simply not performed. Such bias can be avoided if the investigator knows that X_1/X_2 is the critical function of the observables on which to match or weight. However, rarely can we expect investigators to have sufficient theoretical knowledge to unfailingly guess these functional forms. Worse, allowing the investigator to guess at the functional form creates opportunities for selective reporting. Simulations in Section 3.1 further examine this hypothetical example, showing how simple nonlinear functions of the observables can generate large biases when using state-of-the-art matching and weighting estimators, even when bias-adjustment procedures (Abadie and Imbens, 2011) are applied. Kernel balancing mitigates this problem, achieving nearly equal means on X_1/X_2, without the investigator knowing of its importance.

Delaying the technical details until later, the idea behind kernel balancing is straightforward. We first assume that the regression surface for the nontreatment potential outcome (Y_{0i}), conditional on the adjustment covariates, falls in the (reproducing kernel Hilbert) space associated with a choice of kernel. Here, I propose using a Gaussian kernel because the corresponding function space for many smoothly varying outcomes. The practical meaning of this assumption and an interpretation of the resulting function space is provided in this paper. Next, the empirical kernel matrix $\mathbf{K}$, with rows K_i, forms a basis set for the regression function, $\mathbb{E}[Y_{0i}|X_i]$. This amounts simply to a change of bases, from $X_i \in \mathcal{R}^P$ to $K_i \in \mathcal{R}^N$, allowing for highly flexible and complex functions, rather than those simply linear in X_i. Having chosen these bases, kernel balancing finds weights on the control units, such that the weighted average K_i among the control units is approximately equal to the (unweighted) average K_i among the treated units. This is the key step: because the regression surface for Y_{0i} is linear in K_i, achieving (approximately) equal means on K_i ensures (approximately) equal means on Y_{0i} for the treated and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

weighted control groups, without having to fit a model. Weights are chosen by an approximation that minimizes the worst-case bound on the remaining bias due to that approximation. Finally, a simple difference in (weighted) means can then be used to estimate the average treatment effect on the treated (ATT). The remainder of the paper expands upon this logic.

In principal, if one could obtain equal multivariate covariate densities for the treated and control groups, this would nonparametrically and fully adjust for the covariates. In the absence of confounders, this would ensure $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$, which ensures unbiasedness of the difference in means for the ATT, *regardless of the form of* $\mathbb{E}[Y_{0i}|X_i]$. The difficulties with such a "full multivariate density equality" approach are practical: setting aside estimation challenges, this equality cannot even be verified, except when where we have a small number of discrete covariates, each with a small numbers of categorical levels. The simple alternative approach taken here is to first assume that $\mathbb{E}[Y_{0i}|X_i]$ is linear in some set of bases, $\phi(X_i)$. Weights achieving approximately equal means for the treated and control groups on $\phi(X_i)$ then ensure approximately equal means on Y_{0i} between these groups. The analytical framework for kernel balancing elaborates upon this idea, and proposes a particular implementation. Further, while full multivariate density equality is not the aim of kernel balancing, an illuminating connection emerges between that approach and the "balance on kernel-derived bases" approach adopted here. The weights chosen by kernel balancing to achieve mean balance on the chosen bases are exactly those that would equalize the *estimated* empirical multivariate distributions of the covariates for the treated and control groups, *when the same kernel is used for the density estimation.* This reveals a direct link between (1) the assumption one is willing to make about the space of outcome models, and (2) a choice of a smoother, such that the smoothed multivariate covariate distribution is

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

made equal in the treated and control groups.

To briefly place kernel balancing in context, similar to approaches that depend on fitting outcome models (e.g. regression), kernel balancing relies on an assumed outcome model space, although no outcome model ever needs to be fitted. In contrast, as in matching, the dependence on such an outcome model is reduced, because the (weighted) densities of the treated and control groups are made similar in order to compare the two samples on their outcomes, rather than relying on strong modeling assumptions to bridge potentially large gaps between the locations of control and treated observations. On the other hand, kernel balancing also differs from existing matching and weighting approaches. Even when matching methods achieve perfect balance, according to whichever imbalance measures they employ, these balance metrics typically check only for equal means on the covariates, or other moments, as specified by the user. Unfortunately, as in the brief example above, matching discrepancies can give rise to imbalances on unchecked functions of the covariates, leading to biased ATT estimates. Though debiasing methods have been proposed (Abadie and Imbens, 2011), they require a functional form assumption and, thus, ,are not always effective (see Section 3). Kernel balancing also differs from propensity score approaches (Rosenbaum and Rubin, 1983), in that it requires no functional form assumption for the probability of receiving the treatment, given the covariates. This avoids the severe bias that can occur due to possible misspecification of the propensity score (see e.g., Smith and Todd, 2005; Kang and Schafer, 2007). Finally, the method is most similar to weighting or calibration procedures that do not model treatment assignment, but instead achieve balance on covariates (such as Hainmueller, 2012; Zubizarreta, 2015), as well as to analogous survey weighting procedures (Deming and Stephan, 1940). The main contribution of kernel balancing relative to these procedures is that it makes explicit, then

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

weakens, the linear functional form assumption inherent in these weighting and calibration approaches. From an investigator's perspective, the most immediate and practical contribution of kernel balancing is that it provides practioners with a principled and automated answer to the question of what functions of the covariates should be made to have approximately equal means, assuming that the outcome lies in a flexible, smooth space of models.[1]

In what follows, Section 2 provides our analytical framework and develops the method. Section 3 provides a basic simulation, highlighting the dangers inherent in other methods, under reasonable conditions, and demonstrating kernel balancing as a potential solution. Section 4 provides an empirical demonstration of the method's effectiveness in recovering an experimental benchmark from observational data, using the National Supported Work demonstration (LaLonde, 1986). Section 5 presents the implications of this procedure, additional details, and further comparisons to existing matching, weighting, regression, and propensity score approaches. Section 6 concludes the paper. Additional remarks, guidelines, proofs, and empirical examples can be found in the online Supplementary Material.

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 (Splawa-Neyman et al., 1990; Rubin, 1990), let Y_{1i} and Y_{0i} be the treatment and nontreatment potential outcomes, respectively,

[1]Outside the causal inference framework, the same procedure can be used to reweight survey data to match a population of interest, not only on the means of the covariates but on a large space of smooth functions of those covariates.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

for units $i = 1, 2, \ldots, N$, and let $D_i \in \{0, 1\}$ be the treatment assignment for unit i, such that $D_i = 1$ for treated units, and $D_i = 0$ for control units. The observed outcome for each unit is thus $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$. Suppose each unit has a vector of observed covariates, X_i, taking values $x \in \mathcal{X}$, where support $\mathcal{X}$ lies in $\mathbb{R}^P$. For all i, assume that draws of the random variables $\{Y_{1i}, Y_{0i}, X_i, D_i\}$ are taken independently from the common joint density $p(Y_1, Y_0, X, D)$. The set of covariates in $\{X\}$ is assumed to be the set that the investigator must condition upon in order to achieve a causal estimate by ensuring that treatment assignment is ignorable with respect to the potential outcomes, conditionally on the covariates, as assumed under "conditional ignorability,"[2]

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 nontreatment and treatment potential outcomes, repsectively, D_i is tge treatment status, and X_i is a vector of observed pretreatment covariates.

Next, assume $X \in \mathbb{R}^P$ is a set of covariates or characteristics satisfying Assumption 1, and $\phi(X) : \mathbb{R}^P \mapsto \mathbb{R}^Q$, where Q may be (much) larger than P (or N), giving an expanded set of characteristics or features to be used as a set of basis functions.[3] The specific nature of $\phi(\cdot)$ used in kernel balancing will

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

relate to the choice of kernel. For the moment, the key feature of $\phi(\cdot)$ needed is that it is a sufficiently rich, nonlinear expansion such that $\mathbb{E}[Y_{0i}|X_i = x]$ can be well fitted as a linear function of $\phi(x)$:[4]

Assumption 2 (Linearity of Expected Nontreatment Outcome) *We assume that the conditional expectation of Y_{0i} is linear in the expanded features of X_i, $\phi(X_i)$; that is, there exists $\theta \in \mathbb{R}^Q$ and $\phi(\cdot) : \mathbb{R}^P \mapsto \mathbb{R}^Q$, such that*

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

2.2 Population ATT and DIM

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

counterpart, $\mathbb{E}[Y_{0i}|D_i = 0]$. Rewriting this term using Assumption 2, we have

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

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

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

which holds for any θ if

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

The equality in (2) can be interpreted as requiring $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$ in order for the population DIM to be the same as the ATT. Indeed, this mean independence can be used instead of the stronger full independence assumption (Assumption 1) when the average treatment effects are all that are required. Moreover, Equation 3 suggests a natural estimation strategy: owing to the linearity of the assumed function space for $\mathbb{E}[Y_{0i}|X_i]$ (Assumption 2), we obtain $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0]$ whenever $\mathbb{E}[\phi(X_i)|D_i = 0] = \mathbb{E}[\phi(X_i)|D_i = 1]$, regardless of θ, and without need of estimating it. We exploit this fact in the next section.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

2.3 Achieving Mean Balance on $\phi(X_i)$ by Weighting: The Ideal Case

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

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

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

In summary, the ATT is identified by a DIM estimator, modified by the following weights:

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

Thus far, we have established that, in the absence of unobserved confounders (Assumption 1) and the linearity of the conditional expectation of Y_{0i} in $\phi(X_i)$ (Assumption 2), the DIM is equal to the ATT

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

in the population when a $g(x)$ can be found such that the g-weighted expectation of $\phi(X_i)$ among the controls is equal to the unweighted expectation of $\phi(X_i)$ among the treated. We ha,ve chosen bases for the expected nontreatment potential outcome, and ensured equal expectations on each of these bases. Henceforth, we refer to this condition as "mean balance on $\phi(X_i)$." This, in turn, ensures equal expected nontreatment potential outcomes for the treated and control groups, $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_g[Y_{0i}|D_i = 0]$, which ensures that the weighted DIM is equal to the ATT.

2.4 Sample DIM and Weights

We now turn to the sample and the corresponding choice of weights for a plug-in estimator. Let N_0 equal the number of control observations, ,and N_1 be the number of treated observations. We estimate $\mathbb{E}[\phi(X_i)|D_i = 1]$ in Equation 4 using its sample analog, $\frac{1}{N_1}\sum_{i:D_i=1}\phi(X_i)$. For the g-weighted expected nontreatment outcome among the controls, we also replace the expectation with the sample mean, and the "g-weights" with the finite-sample weights $w_1, ..., w_{N_0}$ that solve the sample moment constraints, for all $w_i \geq 0$ and $\sum_i w_i = 1$. Taken together, 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 $w_i \geq 0$, and $\sum w_i = 1$. With w chosen this way (see Section 2.8), we can construct the sample estimator for the DIM, $\widehat{DIM}$. From Equation 5, we replace each expectation in

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

the DIM with the corresponding empirical mean in order to define our estimator,

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

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

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

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

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

Under assumptions of conditional ignorability for the nontreatment 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 the common joint density* $p(X, Y_1, Y_0, D)$.

An alternative derivation begins with a given sample, showing that this weighted difference in means is unbiased for the sample average treatment effect (SATT), which, in turn, is unbiased for the population ATT under random sampling (see the online Supplementary Material, S3). That approach also leads to an analysis of the finite-sample bias under a failure of Assumption 2, that is, 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 ($\mathbb{E}[Y_{0i}|X_i]$) that is not linear in $\phi(X)$ is correlated with the treatment assignment (see the online Supplementary Material, S3.1).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

2.5 Kernels-based Construction of $\phi(\cdot)$

A wide range of basis expansions $\phi(\cdot)$ can be chosen under this estimation framework. Here, rather than choosing $\phi(\cdot)$ directly, we propose doing so implicitly through the choice of kernel, which will generate an N-dimensional vector of features on which equal means can be achieved.

2.5.1 Kernel Notation

For $X_i \in \mathbb{R}^P$, a kernel function, $k(\cdot,\cdot) : \mathbb{R}^P \times \mathbb{R}^P \mapsto \mathbb{R}$, takes as input the covariate vectors from any two observations, and produces a single, real-valued output, interpretable as a measure of similarity between the two vectors. Although numerous kernels can be used in this procedure, for reasons discussed below, we use 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 zero and one, interpretable as a (symmetric) similarity measure, achieving a value close to one when X_i and X_j are most similar, and approaching zero as X_i and X_j become dissimilar. The choice parameter b might be called "scale," because it governs how close X_i and X_j must be, in a Euclidean sense, to be deemed similar (see S12 on the choice of b). It is common to rescale each covariate to have variance 1 prior to computing $k(X_i, X_j)$. This ensures results will be invariant to unit-of-measure decisions. Let the symmetric matrix $\mathbf{K}$ be an N-by-N positive semi-definite (PSD) kernel matrix, with elements $\mathbf{K}_{i,j} = k(X_i, X_j)$. Finally, let the i^{th} row (or column) of $\mathbf{K}$ be written as $K_i = [k(X_i, X_1), k(X_i, X_2), ..., k(X_i, X_N)]$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

2.5.2 Kernel as Inner Product

For any kernel function $k(\cdot,\cdot)$ producing a PSD kernel matrix $\mathbf{K}$, there exists a choice of basis functions $\phi(\cdot)$, such that $\langle\phi(X_i),\phi(X_j)\rangle = k(X_i,X_j)$.[5] 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 corresponds to $\phi(X) = [1, \sqrt{2}X^{(1)}, \sqrt{2}X^{(2)}, X^{(1)}X^{(1)}, \sqrt{2}X^{(1)}X^{(2)}, X^{(2)}X^{(2)}]$; we can confirm that $k(X_i,X_j) = \langle\phi(X_i),\phi(X_j)\rangle$ for this choice of kernel and $\phi(\cdot)$. Using the Gaussian kernel, the corresponding $\phi(X)$ is infinite-dimensional. The function space that is linear in these features can be understood in various ways, as discussed in Section 5.3.

2.6 Mean Balance on K

This section defines mean balance in terms of $\mathbf{K}$ and introduces useful notation. We order the observations such that the N_1 treated units appear first, followed by the N_0 control units. Then, $\mathbf{K}$ can be partitioned into two rectangular matrices,

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

where $\mathbf{K_t}$ is $N_1 \times N$ and $\mathbf{K_c}$ is $N_0 \times N$. The average row of $\mathbf{K}$ among the treated can thus be denoted as $\overline{K_t} = \frac{1}{N_1}\mathbf{K_t}^\top \mathbf{1}_{N_1}$. Kernel balancing seeks weights to ensure that the average row K_i of the treated group is equal to the weighted mean K_i of the control group, which we term "mean balance on $\mathbf{K}$."

[5] This is the result of the equivalence between PSD matrices and the Gram matrices formed by the 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 infinite-dimensional eigenfunctions is given by Mercer's Theorem (Mercer, 1909).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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 $\mathbf{K}$ among the controls.

2.7 Replacing $\phi(X_i)$ with K_i

This section describes how the goal of achieving equal means on $\phi(X_i)$ for the treated and control groups can be replaced by the goal of achieving equal means only on the N-dimensional vectors K_i. Consider fitting $\mathbb{E}[Y_{i0}|X_i]$ using models linear in $\phi(X_i)$, or equivalently, estimating θ in $Y_{0i} = \phi(X_i)^\top \theta + \epsilon_i$ with $\mathbb{E}[\epsilon_i|X_i] = 0$. One might fit such a model using the 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 admit to the representation $\theta = \sum_i c_i \phi(X_i)$. This is proven either by directly seeking to minimize the regularized loss (see, e.g., Hainmueller and Hazlett, 2014) or more generally, by appealing to the Representer Theorem (Kimeldorf and Wahba, 1970). Thus,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

accepting any nonzero degree of regularization, the model will always produce predictions of the form

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

In the case of $\mathbb{E}[Y_{0i}|\phi(X_i)] = K_i c$, we can instead use K_i as bases for the conditional expectation function rather than $\phi(X_i)$; furthermore, these bases do not need to be constructed.

2.8 Choice of Weights: Approximate Balance and Resulting Bias

What remains is to choose the weights w_i to obtain balance on $\mathbf{K}$. Because exact balance on all N dimensions of $\mathbf{K}$ is typically infeasible, we instead seek approximate balance. The approximate balancing approach employed here can be motivated in two different ways: (i) by constructing a worst-case bound on the bias that persists owing to the approximate nature of balance, and minimizing this bound; or (ii) by imagining that we seek balance on a lower-rank approximation of $\mathbf{K}$. While the two are closely related, we take the former as the motivation here. The latter is discussed in the online Supplementary Material, S5.

To derive the worst case bound due to remaining imbalances, by Assumption 2, we can write $\mathbb{E}[Y_{0i}|X_i] = \mathbf{K}c = \mathbf{V}\mathbf{A}\mathbf{V}^\top c = \mathbf{V}d$, where $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{K}$, $\mathbf{A}$ is the matrix whose diagonal contains the eigenvalues of $\mathbf{K}$, and d is a rewritten form of the "coefficients", c, that operate in the eigenvector space, with $d = \mathbf{A}\mathbf{V}^\top c$. Note too that the (Hilbert space) norm of this function is $c^\top\mathbf{K}c = c^\top\mathbf{V}d = d^\top\mathbf{A}^{-1}\mathbf{V}d$. Further, let $\mathbf{V}_1$ be rows of $\mathbf{V}$ corresponding to the treated units, and let

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

$\mathbf{V}_0$ be the rows of $\mathbf{V}$ corresponding to the control units. Then, suppose we choose a vector of weights w_0 on the control units, and w_1 on the treated units. Here, because we target the ATT, every element of w_1 is simply $1/N_1$. The bias of the ATT due to the approximation, denoted as $bias_w$, is then

$$bias_w = \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \tag{8}$$

$$= (w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)d \tag{9}$$

$$= (w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}\mathbf{V}^\top c. \tag{10}$$

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

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

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

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

.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

which, by Cauchy-Schwarz, gives

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

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

The form of this worst-case bound is informative. First, the L_2-norm of the regression function (γ) controls the overall scale of the potential bias. Second, the imbalance on each of the eigenvectors of $\mathbf{K}$ after weighting, $(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)$, enters directly. The contribution of each eigenvector to the bias bound is the product of that eigenvector's imbalance and the square root its eigenvalue. Given this scaling, we choose to achieve near exact balance on the first r eigenvectors, such that the imbalanced eigenvectors are only those with very small eigenvalues, and, thus are of little consequence. Because the matrix $\mathbf{K}$ typically has a few large eigenvalues, followed by many very small ones (if the choice of b is appropriate), it is usually possible to achieve fine balance on eigenvectors that the remaining eigenvalues carry a tiny fraction of the total variation in $\mathbf{K}$.

This is exactly the approach taken here. We chose r so as to minimize *biasbound*, where the norm involved ($\sqrt{\gamma}$) is dropped, because it does not vary across r. Specifically, the weights are found to achieve exact, or nearly exact balance on r dimensions, and the resulting bias bound is computed.[6] Then, r is increased until *biasbound* is minimized. This leaves major imbalances on the relatively inconsequential eigenvectors only, i.e., those with small eigenvalues. While this method minimizes the worst-case bias and appears to be effective in our simulations and applications, future work may fruitfully

[6]The balance on the first r eigenvectors is often exact up to machine precision, but owing to numerical tolerances, small imbalances on these features may persist. The bias bound is written and computed so as to incorporate these residual imbalances, however minor, together with the more substantial imbalances on the remaining eigenvectors.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

propose procedures to minimize $||(w_1^\top \mathbf{V}_1 - w_0^\top \mathbf{V}_0)\mathbf{A}^{1/2}||_2$ using alternative approximate weight-selection methods. The typical behavior of *biasbound* as r increases is illustrated, together with other properties, by the simulations in Section 3.2.

2.8.1 Weight selection, given r

Thus far, we have described the constraints that a set of weights must satisfy (i.e. balance on the first r eigenvectors of $\mathbf{K}$), but not how the weights are chosen. We have great flexiblity in the choice of weights to achieve such constraints and, in particular, a measure of divergence from the uniform weights we wish to keep minimize subject to achieving the balance constraints. The Supplementary Material S1 describes the implementation options consistent with the approach outlined here, including the particular choice implemented in the package `kbal`, which maximizes the entropy measure $\sum_i w_i log(w_i)$, as suggested by Hainmueller (2012). A second choice (also implemented in the `kbal` package) is to use weights that maximize the empirical likelihood, subject to the balance constraints (Owen, 2001). This effectively maximizes $\sum_i log(w_i)$, subject to the constraints. Both methods work well here; choosing between them is beyond the scope of this study. Alternative choices also include the minimum-variance weights described in Zubizarreta (2015), or the nonparametric covariate balancing weights described in Fong et al. (2018).

2.9 Alternative Interpretation: Smoothed Multivariate Balance

The principal motivation for kernel balancing is that it is a reliable and hands-off method for estimating the ATT (or ATC or ATE; see Section 5.4) by obtaining equal means Y_{0i} for the treated and control groups, under reasonable assumptions on $\mathbb{E}[Y_{0i}|X_i]$. However, the use of kernels for the choice of $\phi(X_i)$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

produces a very useful equivalence: kernel balancing using the kernel $k(\cdot,\cdot)$ implies that the multivariate density of the covariates, *as estimated by the same smoothing kernel* $k(\cdot,\cdot)$, will be equal for the treated and control groups, at all covariate locations in the data. Thus, in a finite sample it approximates the goal of "multivariate balance" normally targeted by matching and weighting procedures, but only insofar as those densities are well estimated using that choice of kernel.

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

Proposition 1 (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 using the kernel* $k(\cdot,\cdot)$ *of bandwidth* b*. 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 in* $\mathcal{X}$ *where an observation is found.*

The proof of Proposition 1 is given in the Supplement Material S7. Here, I briefly describe the intuition behind this result, because it leads to further insights and tools. First, the typical Parzen–Rosenblatt window approach estimates a density function according to:

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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

The link between obtaining mean balance on Y_{0i} and obtaining multivariate density balance emerges from the fact that both involve the superposition of rescaled kernels placed over each observation. For a sample consisting of $X_1, \ldots, X_N$, construct the kernel matrix $\mathbf{K}$ using the Gaussian kernel, and right-multiply it by a column vector, $\frac{1}{N\sqrt{4\pi b}}$. This produces values numerically equal to first constructing such an estimator based on all observations represented in the columns of $\mathbf{K}$, and then evaluating the resulting density estimates *at all 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 an N_1-vector of ones, 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 can analogously rewrite the estimated density of the treated as $\frac{1}{N_1\sqrt{4\pi b}}\sum_i K_i$ and the weighted estimated density of the controls as $\frac{1}{\sqrt{4\pi b}}\sum_i K_i w_i$. Setting these equal to each other gives $\frac{1}{N_1}\sum_i K_i = \sum_i K_i w_i$, which is the same condition (mean balance on K_i, Definition 1) pursued by kernel balancing. This reveals the deep connection between (i) the assumption we make on the space of models for the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

outcome, and (ii) the choice of smoother, for which estimated multivariate density balance is achieved.

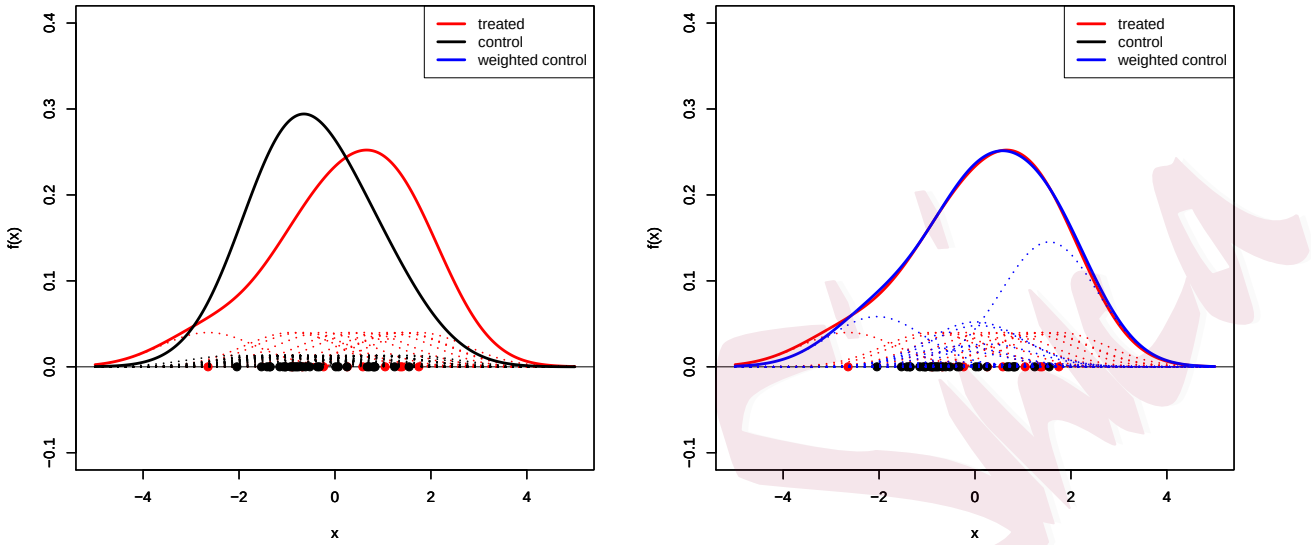
Of practical relevance to investigators, this suggests a measure of imbalance relating to the difference between the kernel-estimated distribution of the covariates in the treated and for the control groups, both before and after weighting. We first consider the goal of achieving mean balance on K_i. To minimize the smoothed multivariate imbalance under this kernel, we minimize a p-norm proportional to $||\overline{K_t} - \sum_{i:D=0} w_i K_i||_p$. On the other hand, we can calculate a norm over the "difference in heights" between the implied density estimates for the treated group and control group, at every observation's covariate location. This is given by $\frac{1}{2}||\hat{p}_{D=1}(\mathbf{X}) - \hat{p}_{w,D=0}(\mathbf{X})||_p$. Fortunately, we need not choose, because 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 the Supplementary Material S2 for details.

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

Figure 1 illustrates the density-equalizing property of the kernel balancing weights for a one-dimensional problem. This density equalizing view connects kernel balancing more directly to other approaches, such as matching, but note that it is mean balance in Y_{0i}, achieved through mean balance on a suitable set of bases (K_i), that kernel balancing targets, that is essential for unbiasedness of the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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

ATT, and that gives rise to the bias bound and other analytical results.

3 Simulation Examples and Evidence

3.1 An Illustration: Imbalance on a Ratio

Building on the simple example given in Section 1, this simulation highlights the practical challenges of existing methods and demonstrates the effectiveness of kernel balancing against these challenges. For realism, suppose we are interested in the question of whether peacekeeping missions deployed after civil wars are effective in lengthening the duration of peace (*peace years*) after the war's conclusion (e.g., Fortna, 2004; Doyle and Sambanis, 2000). However, within the set of civil war cases constituting our

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

sample, the "treatment" — peacekeeping missions (*peacekeeping*) — is not randomly assigned. Rather, missions are more likely to be deployed in certain situations, which may differ systematically in their expected *peace years*, even in the absence of a peacekeeping mission. Suppose the investigator therefore collects four pre-treatment covariates: the duration of the preceding war (*war duration*), the number of fatalities ($fatalities$), the 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 experienced peacekeeping missions.

Further, suppose there are no unobserved confounders, and that peacekeeping missions are deployed only on the basis of these observables. Specifically, consider a conflict's *intensity*, given by $\frac{fatalities}{war\,duration}$, and suppose that missions are more likely to be deployed in higher-intensity conflicts,

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

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

Further, suppose the outcome of interest, *peace years*, is also a function of *intensity*, with more intense conflicts leading to longer duration of *peace years* on average:

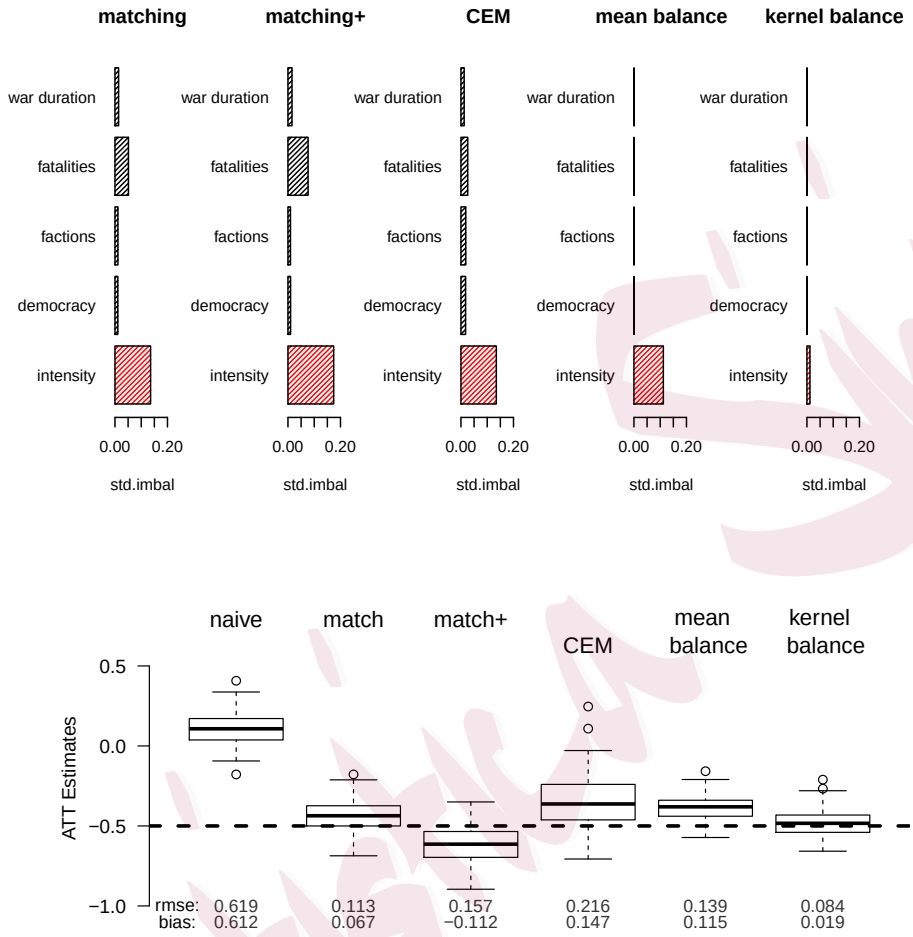
$$peace\,years = 5 + 2\frac{intensity}{5000} - (0.5)peacekeeping_i + \epsilon_i,$$

where and ϵ_i is an error term drawn from $N(0, 4)$. This arrangement generates a fixed treatment effect

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

of -0.5 years.[7]

Figure 2: Simulation: Imbalance on a Ratio



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

How well do existing techniques achieve equal means for the treated and control groups ("mean

[7]Such a confounding scenario may occur if, 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

balance"), both on the original four covariates and on *intensity*, a (nonlinear) function of the observables? In Figure 2, the top panel shows the covariate imbalance on the horizontal axis (the standardized difference in means between treated and control), for each of the covariates and the key function of the covariates, *intensity*. All results are taken over 500 simulations, with the same data-generating process and $N = 500$. First, *matching* (simple Mahalanobis distance matching with replacement) leaves a substantial imbalance on *war duration*, and more troubling, on *intensity*. 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 results for *matching+* show that even this approach would not generate balance on *intensity*. In fact, balance on both *war duration* and *intensity* has worsened. Next, coarsened exact matching (*CEM*) coarsens the variables so that exact matching on the resulting data is possible (Iacus et al., 2011). However, this does not solve the problem: imbalances remain on the original, uncoarsened variables. Fourth, (*mean balance*) employs entropy balancing (Hainmueller, 2012) to achieve equal means in the original covariates. As expected, this produces excellent balance on the original covariates, but only a modest improvement in balance on *intensity*. Finally, *kernel balance* achieves vastly improved balance on *intensity*.

These imbalances are worrying, because they indicate a failure to condition on the covariates as intended. Because an imbalanced covariate directly influences the potential outcomes, this imbalance leads to biased ATT estimates. To show this, the ATT estimate for each method is shown in the bottom panel of Figure 2. Large biases occur for each estimator, with the exception of kernel balancing. Note that for both *matching* and *matching+*, bias adjustment Abadie and Imbens (2011) is employed in

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

an effort to make up for matching discrepancies on the observed covariates. However, this does not account for the nonlinear effects of the observables on the potential outcomes. Kernel balancing shows the lowest bias among the methods attempted. Its advantages in terms of the RMSE are more modest, but it still has an RMSE 22% lower than that of the next best estimator, *mean balance*.

Although kernel balancing is largely automated, given a choice of Gaussian kernel, we still need to choose the bandwidth parameter, b. Section S12 describes the substantive meaning of this parameter, but it is useful to examine the sensitivity of results to choices of b. Figure 4 in the online Supplementary Material 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 S12 for discussion of this default value).

This illustration demonstrates the ease with which existing methods may fail: when a confounder is a nonlinear function of two observed covariates — even a simple ratio — existing matching and weighting methods risk large biases. An investigator's theoretical knowledge is rarely sufficient to which functions of the observables may impact the outcome. Kernel balancing provides a principled approach for choosing the functions of covariates on which to achieve balance to ensure unbiased estimation in a wide range of plausible scenarios, specifically those where the non-treatment potential outcome is a smooth function of X_i. An illustration of the effectiveness of the worst-case bound in bounding biases under this simulation is given in the online Supplementary Material S4.

3.2 Behavior of Bias Bound and L_1 imbalance across r

In another simulation, we illustrate the behavior of the L_1 imbalance and the *biasbound* across levels of r, while examining whether minimizing these quantities is an effective way to minimize the imbalance

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

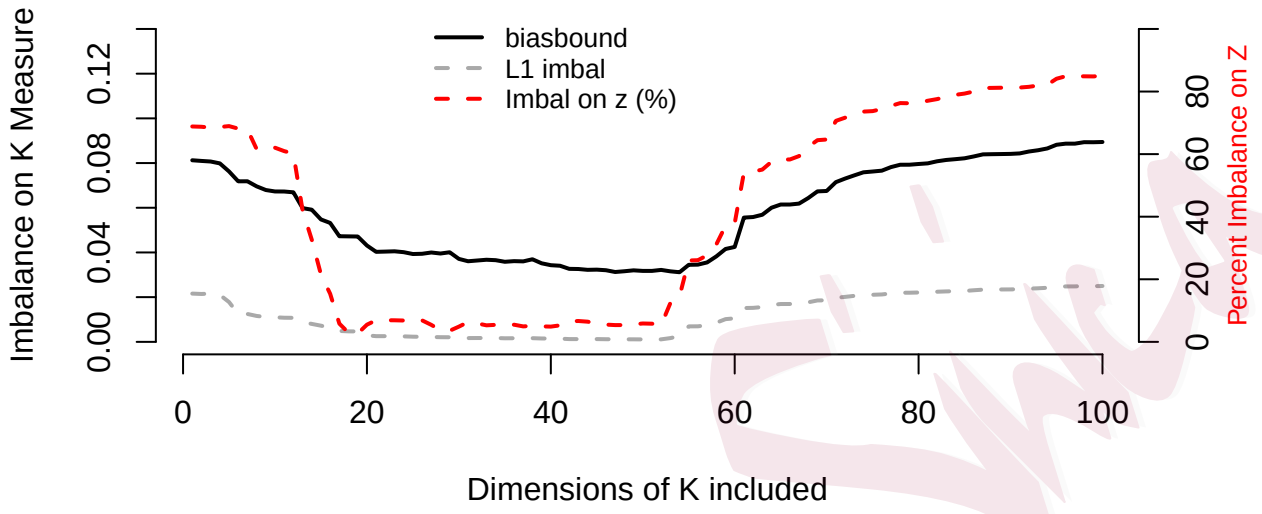
on important, but unknown functions of the covariates. Let $x_{1i}, ..., x_{5i}$ be covariate data, each drawn $N(0,1)$, for $i \in 1, ..., 500$, and let $z_i = \sqrt{x_{1i}^2 + x_{2i}^2}$. This function impacts the treatment assignment, with the probability of treatment given by $logit^{-1}(z_i - 2)$, producing approximately two control units for each treated unit. In Figure 3, the number of factors of $\mathbf{K}$ retained for purposes of balancing (given by r) is increased from 1 up to 100. The bias bound shown here is not scaled – i.e. it is computed as if $\gamma = 1$ to illustrate how it changes across r given a constant choice of γ. As expected, the bias bound, the L_1 score, and the imbalance on z (as a percentage of the original imbalance) after weighting improve rapidly as r increases, with the most important eigenvectors coming into balance. It then plateaus, and eventually worsens beyond some choice of r. Most importantly, while the balance on z is unknown to the investigator, the bias bound and L_1 are observable, and improvements in balance on z are strongly correlated with improvements in the other two. Accordingly, selecting r to minimize the bias bound appears to be a viable strategy for selecting the value that also minimizes imbalance on unseen functions of the data. As expected, the bias bound and L_1 are very similar, up to a scaling factor; in fact, all three quantities in Figure 3 correlate with each other above 0.96. Note too that there is a wide range of r values (approximately 20 to 50) that produce similar levels of imbalance, making the exact choice less critical.

4 Example: National Supported Work Demonstration

It is useful to know whether kernel balancing accurately recovers average treatment effects in observational data under conditions in which an experimental benchmark is available for comparison. This can be approximated using the approach and data of LaLonde (1986) and Dehejia and Wahba (1999), now a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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



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

routine benchmark for matching and weighting approaches in disciplines as diverse as statistics, econometrics, political science, psychology, and epidemiology (see, e.g., Diamond and Sekhon, 2005; Iacus et al., 2011; Hainmueller, 2012; McCaffrey et al., 2004; Little and Rubin, 2000). The aim of these studies is to recover an experimental estimate of the effect of a job training program, the National Supported Work (NSW) program. Following LaLonde (1986), the treated sample from the experimental study is compared to a control sample drawn from a separate, observational sample. Methods of adjustment are tested to determine whether 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 these data and various subsets that have been drawn from it. Here, I use

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

185 treated units from NSW, originally selected by Dehejia and Wahba (1999), for the treated sample. For this group, the experimental estimate of the ATT is \$1794, which gives us a benchmark. For the observational version of the study, we keep these treated units, but draw the control sample from the Panel Study of Income Dynamics (PSID-1), containing 2490 individuals.

The pretreatment covariates are age, years of education, an indicator for no high school degree, real earnings in 1974, real earnings in 1975, indicators for zero income (taken to mean unemployment) in 1974 and 1975, and a series of demographic indicator variables: black, hispanic, and married. As found by Dehejia and Wahba (1999), propensity score matching can be effective in recovering reasonable estimates of the ATT, but these results are highly sensitive to the specification choice used for propensity score estimation (Smith and Todd, 2001). Diamond and Sekhon (2005) use genetic matching to estimate the treatment effect 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 squared terms and two-way interactions, as well as constructed indicators for zero income in 1974 and 1975. Similarly, entropy balancing (Hainmueller, 2012) has been shown to recover good estimates using a similar setup, using a control dataset based on the Current Population Survey (CPS-1), employing all pairwise interactions and squared terms for continuous variables, amounting to 52 covariates. The general supposition of kernel balancing, however, is that investigators would not typically know (or be expected to know) that what nonlinear transformations are required to obtain a good estimate.

In this re-analysis, three estimation approaches are compared, with three specifications attempted for each. The first procedure is a simple linear regression (OLS), which is not an effective competitor, but serves to show that assuming a simple outcome model can produce highly problematic results. Second,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

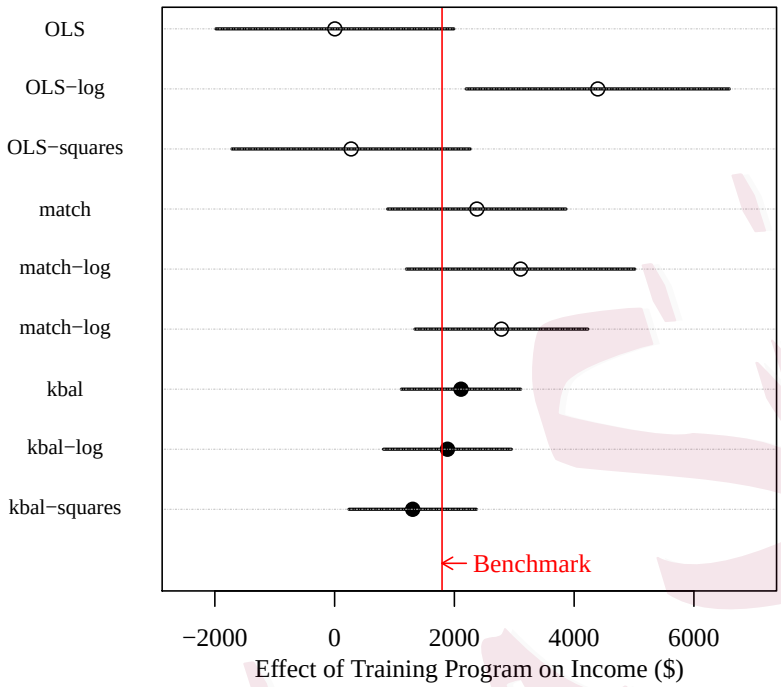
Mahalanobis distance matching (*match*) is employed, with bias adjustment. Third, kernel balancing (*kbal*) is used, with b set to the default value of P (the number of covariates). For comparability, all three approaches use simple standard errors that take the weights as fixed.

For each method, three specifications are attempted, chosen on the grounds that they are reasonable choices we might expect investigators to make and justify in their analyses. First, we include the "standard" set of 10 covariates described above. Second, an investigator might reasonably propose that log income is a better choice than raw income for determining who should be considered similar and, thus, be matched together. Therefore, incomes in 1974 and 1975 are replaced with their logs (plus one). Third, a thoughtful investigator may be concerned about flexible functional forms and try an expanded set of covariates including the 10 standard covariates, plus the squared terms for the three that are continuous. Note that all three of these approaches appear to be arguably justifiable.

Figure 4 shows that OLS estimates do poorly and vary widely by specification. This reflects the large differences between the distributions of treated and control units in the covariate space, and that using a linear model to account for these imbalances fails. Matching performs much better, though remains somewhat specification-dependent, with its best estimate (*match-squares*) falling within \$581 of the benchmark. Finally, kernel balancing (*kbal*) performs well across all three specifications, with no estimate more than \$490 from the benchmark, and the average estimate only \$27 off. Whether constructing additional squared terms or taking the log of income, the space of functions represented in the span of $\mathbf{K}$ is large and flexible, so the resulting solutions change little. The resulting ATT estimates are also very stable to the choice of b, once it exceeds a minimum value (see Figure 5 in the online Supplementary Material).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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



Re-analysis of Dehejia and Wahba (1999), estimating the effect of a job training program on income. Three procedures are used: linear regression (*OLS*), Mahalanobis distance matching (*match*) with bias adjustment, and kernel balancing (*kbal*). For each, three sets of covariates are attempted: the standard set of 10 covariates described in the text, a version replacing income in 1974 and 1975 with log income (*log*), and an expanded set (*squares*) including the 10 standard covariates plus squares of the three continuous variables. The experimental benchmark of $1794 is indicated by the vertical line. While both *match* and *kbal* produce reasonable results, *kbal* results are closest to the benchmark, showing the least sensitivity to the specification.

Examining additional outputs from the kernel balancing procedure, we further see the that the initial sample was badly imbalanced, with an L_1 distance of 78%. Fortunately, weights can be found to eliminate much of this, because the L_1 distance drops to 4.7% after weighting. At the solution achieved by kernel balancing using the default value of $b = 10$, 90% of the weight for controls is taken from just 66 units.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

5 Discussion

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

5.1 Relation to Existing Approaches

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

5.1.1 Matching

Under conditional ignorability (as in Assumption 1), sub-classification and exact matching estimators for the average treatment effect X implement conditioning very literally: take difference-in-means estimates within each stratum of X, then average these over the empirical distribution of X for the treated. However, conditioning on X in this way is impossible when X is continuous or contains indicators for many categories. Matching approaches (e.g., Rubin, 1973) mimic this conditioning, taking each treated unit in turn, finding the nearest one or more control units, and retaining only these control units in the sample. 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; that is, when the distribution of X for the control units becomes the same as the distribution for the treated units. The nonparametric nature of matching is

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

appealing as a multivariate balancing technique, but its accuracy is limited by the problem of matching discrepancies. Specifically, in a given pairing, the treated unit may be systematically different on X than the control unit(s) it is paired with when exact matches cannot be found. Thus, the conditioning on X can remain incomplete, and the distribution of X for the treated and controls will not be made identical. The resulting bias in (S)ATT estimates dissipates very slowly as N increases, so that the resulting estimates are not generally $\sqrt{N}$-consistent (Abadie and Imbens, 2006). To minimize the bias due to remaining matching discrepancies, investigators are sometimes instructed to attempt different matching specifications and procedures until they achieve satisfactory multivariate balance (see, e.g., Stuart, 2010). However in practice, tests for this balance are usually limited to univariate tests that compare the marginal distribution of each covariate under treatment and control. As the simulation example in Section 3.1 illustrates, many matching approaches can thus fail to obtain a sufficient similarity of distributions, even when investigators attempt to match on higher-order terms. Relatedly, a class of methods referred to as *optimal* matching minimizes a global measure of distance between the distributions of treated and control units. Kallus (2016) considers a generalization of optimal matching methods, considering a bias-variance tradeoff and choosing the point that minimizes the worst-case conditional mean squared error. The study proposed kernel optimal matching, which solves a minimization problem involving a Gaussian or other kernel representation of the data in the minimization objective. This proves to have many useful properties, especially when paired with an outcome regression model, suggesting another route by which kernels may be useful for estimating treatment effects.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

5.1.2 Covariate Balancing Weights

Covariate balancing weighting approaches use probability-like weights on the control units to achieve a set of prescribed moment conditions on the distribution of the covariates (e.g., univariate means and variances). Examples from the causal inference literature include entropy balancing (Hainmueller, 2012) and the covariate balancing propensity score (Imai and Ratkovic, 2014; Fong et al., 2018), as well as relate procedures noted in earlier work on survey sampling, such as raking (Kalton, 1983). Once these moment conditions are satisfied, it is assumed that the multivariate densities for the treated and control are alike enough to complete the adjustment. These weights can be used in a difference-in-means estimation or other procedure. The advantage 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 disadvantage is that it sacrifices the nonparametric quality of matching, providing balance only on enumerated moments. In general it is 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 (nontreatment) outcome. Kernel balancing can be understood as an extension of these covariate balancing weighting methods that chooses moments to ensure balance on by constructing bases that span a flexible space for the outcome model.

5.1.3 Propensity Score Weighting

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

For the purposes of ATT estimation, the stabilized inverse propensity score weights applied only to the control units would be $w_{IPW} = \frac{p(D_i)}{p(D_i|X_i)} \frac{1-p(D_i|X_i)}{1-p(D_i)}$. The Supplement Material S10 shows how these weights are derived to change the distribution of the controls to match that of the treated during the ATT estimation. Moreover, these weights can be rewritten using Bayes' rule as a ratio of class densities for the treated and controls,

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

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

5.1.4 Comparison with 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 the treatment D_i. Combining the power and flexibility of machine learning or high-dimensional models with an outcome model that efficiently and unbiasedly returns estimates of causal effects remains an active area of research. Regularized regression models are employed

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

to accomodate high-dimensional covariates. However the shrinkage imposed by these models leads to substantial bias and poor inferential properties (Belloni et al., 2014). A series of doubly-robust or debiasing methods utilizing a (consistent) estimator of the propensity-score to adjust these models have been proposed, following Robins et al. (1994) (see, e.g., Van der Laan and Rose, 2011; Farrell, 2015; Belloni et al., 2017; Chernozhukov et al., 2017). Further recent efforts have sought to avoid the requirement of a consistently estimated propensity score model to make such adjustments. For example, Ratkovic and Tingley (2017) propose a Bayesian sparse model for variable selection that, combined with feature expansions such as a tensor-spline, performs well and can be easily extended to other approaches. Athey et al. (2016) effectively combine the outcome model approach with weights that seek covariate balance, by using the covariate balancing weights to re-weight residuals from a linear outcome model. Like these methods, kernel balancing adopts insights from machine learning, relying on the properties of kernels. However, unlike other studies that import machine learning methods, it does not use them to solve a classification or regression problem, such as fitting the outcome or a propensity score. Rather, it uses kernels to establish a high-dimensional choice of bases, which tell us what functions of the covariates must be balanced.

Two important distinctions can be made between assuming an outcome model for the purpose of choosing "what to balance on," as done here, and fitting an outcome model. First, kernel balancing works regardless of *which* function in the function space is the correct one (i.e., the value of θ or $\mathbf{c}$). We do not need to rely on the accuracy of any estimate of these coefficients. We require only that such a model exists, and even then, violations of the model are bias-inducing only in certain cases (see Supplementary Material S3.1). Second, employing a weighting approach justified by a choice of

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

outcome models is not equivalent to using the outcome model alone, because when estimating the ATT, the former changes the distribution of the control group to be more similar to that of the treated, prior to estimation of an effect. This "pre-processing" approach reduces model-dependency, avoiding strong modeling assumptions to bridge the gap between treated and control units that may lie far apart in the covariate space (see, e.g., Ho et al., 2007 for analogous arguments in the matching literature). That said, future work could usefully combine the kernel balancing technique with a suitable outcome model in an augmented regression or doubly-robust procedure.

5.2 Uncertainty Estimation

In most contexts, investigators require a measure of uncertainty, such as a standard error or confidence interval, around their effect estimates. With matching estimators, one approach is to ignore the uncertainty due to the matching procedure itself. For example, Ho et al. (2007) argue that because the variance estimators for parametric models typically take the data as fixed anyway, when data are pre-processed by a matching procedure, the matched data set can be taken as fixed for subsequent analyses. Thus, the variance can be estimated for parametric outcome models on the matched data in the usual way: constructing weights that reflect the selection of matched control-units, then estimating the outcome model with these weights to obtain the associated standard errors. Similarly, weighting estimators such as entropy balancing may also take this preprocessing view and treat the resulting weights as fixed (Hainmueller, 2012) when 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 owing to the "extreme nonsmoothness" of matching. Abadie

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

and Imbens (2006) develop asymptotic standard errors that account for uncertainty in the matching procedure. Others have argued that an m-out-of-n bootstrap may be appropriate (see Politis and Romano, 1994). One benefit of kernel balancing and other weighting methods is that, because the weights are continuous and observations are not wholly dropped, as in matching, the simple bootstrap may be valid. However, the development of more computationally attractive alternatives remains an area for ongoing research.

5.3 Gaussian Kernel and Intuition for $\phi(X_i)$

One reason to use the Gaussian kernel is that it is the workhorse kernel in machine learning regression and classification tasks, and so the feature space it implies is likely to be an appropriate one. Though kernel balancing does not actually fit an outcome model here, the function space invoked, $\phi(X_i)^\top\theta$, is the same as that in which kernel methods, such as kernelized regression, support vector machines with kernels, and Gaussian processes, operate. Moreover, the Gaussian kernel has the universal representation property: as $N \to \infty$, $\phi(X)^\top\theta$ it encompasses any continuous function of X (Micchelli et al., 2006). While asymptotically appealing, this universality as N approaches ∞ is less reassuring in finite samples. Nevertheless, 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 typically expect that small changes in X_i should lead to small changes in Y_{0i}.

One approach to better understanding this function space is to analyze the features, $\phi(\cdot)$, consistent with the Gaussian kernel, that is, those for which $\langle\phi(X_i), \phi(X_j)\rangle = k(X_i, X_j)$ holds true. Because the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

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. A valid choice for $\phi(X)$ in the case of the Gaussian kernel (with one-dimensional X) is $\left\{\sqrt{\frac{2^d}{d!}}exp(-X_i^2)(X_i)^d\right\}$ for $d = 0, 1, ..., \infty$. The Supplementary Material S8 describes this in greater detail, but fortunately a more intuitive understanding of this function space is available. As shown above, the functions linear in $\phi(X_i)$ are also those linear in K_i. Accordingly, $k(X_i, \cdot)$ is sometimes called the "canonical feature mapping" corresponding to $\phi(x)$, (e.g., Minh et al., 2006). Because $k(X_i, x)$ evaluates at x the height of a Gaussian that had been centered at X_i, this function space is that which can be built a by superposition of Gaussians placed over each observation and arbitrarily rescaled. That is, in the original covariates space $\mathbb{R}^P$, suppose we place a p-dimensional Gaussian kernel over each observation in the data set, rescale each of these by a scalar c_i, and then sum these rescaled Gaussians to form a single surface. By varying the values of c_i, a wide variety of smooth functions can be formed, approximating many nonlinear functions of the covariates. This view is described and illustrated at length in Hainmueller and Hazlett (2014), where the same function space is used to model smooth, highly nonlinear functions.

Another key question in determining what kernel to use is the choice of bandwidth, b. A useful default value for b is given by the column rank of X (see the Supplementary Material S12). An easy and transparent guideline for investigators is to show results using a range of choices for b, starting from one half of the default value ($\dim(X_i)$) up to several times that value. Further details on the choice of b and its implications are discussed in the Supplementary Material S12. The stability of our results over choices of b in both the simulation and the applied example are illustrated in Supplementary Material S13.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

5.4 Other Quantities: ATE, ATC

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

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

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

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 the control to the same location and (more importantly) to have the same expectation o.f Y_{0i}. Thus, we 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 observations, treated and control alike. The `KBAL` package estimates the ATT by default, but optionally estimates the ATC and ATE as well.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

6 Conclusion

In order to reliably infer causal quantities from observational data, the primary challenge is often ensuring that we observe a set of variables that are collectively sufficient for achieving conditional ignorability. However, even then, performing the required conditioning on observables, particularly with multiple, continuous covariates, remains nontrivial. Matching, covariate balancing weighting, and propensity score weighting each seek to make the multivariate distribution of covariates for the untreated more similar to that of the treated. If any function of the observables that systematically influences the nontreatment outcome persists in having a different mean for the treated and controls, the resulting estimates may be biased. Unfortunately, the investigator is not usually aware of all the functions of the covariates that may influence the outcome, making it difficult to guard against this possibility. As illustrated here, when even a simple nonlinear function of observables is confounding, existing methods can fail to complete the desired adjustment.

Fortunately, the unbiased estimation of the ATT requires only that the expected nontreatment potential outcome is equal in the treated and control groups after adjustment. Kernel balancing seeks to achieve this by first assuming that $\mathbb{E}[Y_{0i}|X]$ falls in a large space of smooth functions, which is, in turn, linear in the columns of the kernel matrix, $\mathbf{K}$, rather than the original covariates, X. It finds weights on the controls to make the weighted average row of $\mathbf{K}$ for the controls approximately equal to the average row of $\mathbf{K}$ for the treated. This ensures that the expected nontreatment outcome is approximately equal in the two groups. Bias owing to the approximate nature of the weights can be bounded, and the weights are chosen using a method that minimizes this worst-case bias. An alternative interpretation of the procedure is that kernel balancing implies that a particular kernel-based smoother

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0555

for the multivariate densities is approximately equal for the treated and control groups, as evaluated at every observation.

Numerous extensions remain for future work. First, $\mathbf{K}$ has dimensionality $N \times N$, which becomes unwieldy as N grows large, posing a practical limit of tens of thousands of observations. Second, while the bootstrap may provide confidence intervals that include uncertainty due to weight selection, further work is needed, particularly on approximations that may not be as computationally burdensome when N is large. Finally, improvements may be possible on a number of implementation details, such as the choice of the kernel and its parameters, the optimization procedure and choice of methods for achieving approximate balance, and understanding the potential for bias owing to imperfect balance, using metrics that are less extreme than the worst-case bound. An implementation of this procedure using the choices described here is available in the `R` package `KBAL`.

7 Supplementary Material

The online Supplementary Material provides proofs, remarks, additional simulations and illustrations, and additional empirical applications.

References

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.