2016

A kernel-based metric for balance assessment

Yeying Zhu, *University of Waterloo*
Jennifer S Williams, *The Pennsylvania State University*
Debashis Ghosh

1 Introduction

In many scientific and medical studies, there has been a lot of interest to determine the causal effects associated with an intervention in an observational study setting. While the "gold-standard" approach would be randomization to the intervention, in many situations, this cannot be done because of logistic, economic, and/or ethical constraints. While potentially controversial from a scientific point of view, there has been a renewed interest from the statistical perspective in terms of the analysis of data from observational studies.

A central model for the formulation of causal effects has been the potential outcomes framework (Neyman, 1923, Rubin, 1974). Within this setup, an important quantity to facilitate causal inference has been the propensity score (Rosenbaum and Rubin, 1983), defined as the probability of receiving the treatment given a set of measured covariates. Using the propensity score, causal inference proceeds in two stages. At the first stage, the propensity score is modelled as a function of predictor variables. The second stage involves causal effect estimation in which the propensities are used for adjustment. Methods following this two-stage idea include inverse probability weighting, matching and subclassification.

A key issue in performing causal inference is the assessment of balance. Roughly speaking, balance means that the distribution of confounders between the treatment and control groups are equal. From Rosenbaum and Rubin (1983), the assumptions of strongly ignorable treatment assignment (defined in §2), in conjunction with the definition of the propensity score, imply that adjustment on the propensity score will theoretically achieve balance. While this holds in theory, the practical assessment of balance remains an important issue. Methods for balance diagnostics have been proposed by Ho, Imai, King, and Stuart (2007) and Sekhon (2011). Austin and Stuart (2015) provided a comprehensive review of different quantitative and qualitative balance metrics based on the inverse probability weighting procedure.

Several recent proposals have focused on the modeling of propensity scores or inverse probability weights via balancing covariates. The underlying idea with this approach is that by achieving balance in the covariates, the bias due to measured confounders can be reduced (Harder, Stuart, and Anthony, 2010). Examples include the GBM (McCaffrey, Ridgeway, and Morral, 2004), entropy balancing (Hainmueller, 2011), CBPS (Imai and Ratkovic, 2014) and kernel balancing (Hazlett, 2015). In some of these approaches, a balance statistic, such as the average standardized absolute mean difference , c-statistic or Kolmogorov-Smirnov statistic is optimized in a certain way. Consequently, even when the propensity score model is incorrect, there is enough overlap in the covariates to draw reliable causal inferences. In other words, the causal estimates are robust to model misspecification.

In Hazlett (2015), the author shows if the covariates have linear effects on the potential outcomes, the sample mean difference in the observed outcome is unbiased for the causal effect when one achieves balance in the means of the covariates. However, the linearity assumption could be a very strict assumption and when the covariates do not have linear effects on the outcome, achieving balance in the finite moments of the covariates may not be enough. In Section 2, we review the potential outcomes framework and point out the concept of achieving balance is multivariate in nature and ideally involves guarantees that the distributions of the confounders between the two treatment groups are equal. Using reproducing kernel Hilbert space (RKHS) methodology, we propose a new metric for assessing covariate balance in Section 3, called kernel distance. Section 4 features simulation studies evaluating the proposed kernel distance against several other balance measurements from the literature. In Section 5, we apply the methodology to the Early Dieting in Girls study, in which we aim to draw causal inference about mothers' weight concern on daughters' early dieting behavior using matching. We conclude with some discussion in Section 6.

2 Background and Preliminaries

2.1 Data Structures and Causal Estimands

Let the data be represented as (Y_i, T_i, Z_i), $i = 1, \ldots, n$, a random sample from the triple (Y, T, Z), where Y denotes the response of interest, T denotes the treatment group, and Z is a p-dimensional vector of covariates. We assume that T takes the values $\{0, 1\}$.

We briefly review the potential outcomes framework (Rubin, 1974, Holland, 1986) in order to define the target estimands that will be of interest. We define counterfactuals $(Y(0), Y(1))$ for all n subjects, and the observed response is related to the counterfactuals as $Y \equiv (1 - T)Y(0) + TY(1)$. Causal effects are defined as within-individual contrasts based on the counterfactuals. For example, given $(Y_i(0), Y_i(1))$, $i = 1, \ldots, n$, we can define the average treatment effect:

$$\text{ATE} = E[Y(1) - Y(0)]. \tag{1}$$

Another quantity we will consider in this article is the average treatment effect among the treated (ATT), whose population parameter is defined as

$$\text{ATT} = E[Y(1) - Y(0) | T = 1]. \tag{2}$$

We note in passing that (1), when defined for the subpopulation with $T = 1$, gives an alternative formulation to (2).

An important assumption for valid causal inference is the strongly ignorable treatment assumption (Rosenbaum and Rubin, 1983):

$$T \perp \{Y(0), Y(1)\} | Z. \tag{3}$$

Assumption (3) means that treatment assignment is conditionally independent of the set of potential outcomes given covariates.

To estimate causal effects in observational studies, Rosenbaum and Rubin (1983) proposed the use of the propensity score, which is defined as

$$e(Z) = P(T = 1|Z). \tag{4}$$

In words, (4) represents the probability of receiving treatment as a function of covariates. Given the treatment ignorability assumption in (3), it also follows by Theorem 3 of Rosenbaum and Rubin (1983) that treatment is strongly ignorable given the propensity score, i.e.

$$T \perp \{Y(0), Y(1)\} | e(Z).$$

Based on the collection of these assumptions, we can view the process of estimating causal effects as a two-step process. At the first stage, the analyst models the propensity score as a function of the available covariates. The second stage involves estimation of the causal effect by modelling the effect of T on Y with adjustment using the estimated propensity scores from the first step.

2.2 Balance using Probability Metrics

In this section, we wish to study covariate balance from the viewpoint of comparing the multivariate distributions of $Z|T = 1$ and $Z|T = 0$. Denote these distributions as F_1 and F_0 with corresponding dominating probability measures P_1 and P_0. We now review the concept of probability metrics, an overview for which can be found in Zolotarev (1983):

Definition 1. Let X, Y, Z denote random variables that are defined on a common probability space $\mathscr{P}$. Then a probability metric is a mapping $d : \mathscr{P} \times \mathscr{P} \to [0, \infty)$ that satisfies the following properties:

(PM1) If $P(X = Y) = 1$, then $d(X, Y) = 0$.
(PM2) $d(X, Y) = d(Y, X)$.
(PM3) $d(X, Z) \leq d(X, Y) + d(Y, Z)$.

These properties mostly have natural analogues to distances when applied to real numbers. Property (PM3) is the triangle inequality, while (PM2) represents symmetry of the probability metric in its arguments. Property (PM1), strictly speaking, is not a property of a distance metric but rather that of a semimetric. It states that if two distributions are equal in law, then their probability metric value will be zero. Finally, we point that for (PM3), if the left-side is infinity, then one of the terms on the right-hand side of the inequality must also be infinitely for the triangle inequality to hold. Zolotarev (1983) and Rachev, Klebanov, Stoyanov, and Fabozzi (2013) give comprehensive overviews on probability metrics. In an abuse of notation in this article, we will use (X,Y), (F_1,F_0) and (P_1,P_0) as the arguments for the probability metric. Given the definition of probability metric, we can now define distributional covariate balance (DCB) as $d(F_1,F_0) = 0$ where F_1 and F_0 are the multivariate distributions of $Z|T = 1$ and $Z|T = 0$. We see that in effect, satisfying this definition guarantees the covariate overlap needed for proper causal inference. We point out that DCB enforces equality on the *joint* distribution of confounders given treatment groups. Thus, it is stronger than the CBPS proposal of Imai and Ratkovic (2014), which is implemented requiring only the first or second moments of the confounders to be equal across treatment groups. While DCB seems like a desirable property to obtain, its implementation on real datasets seems to not be straightforward. A useful device to aid in this is Reproducing Kernel Hilbert Spaces (RKHS), which we now discuss.

2.3 Reproducing Kernel Hilbert Spaces

More comprehensive overviews on RKHS can be found in Wahba (1990), Berlinet and Thomas-Agnan (2011) and Steinwart, Hush, and Scovel (2006). As the name suggests, RKHS is a Hilbert space $\mathscr{F}$ with inner product $< \cdot,\cdot >_{\mathscr{F}}$ whose elements are functions $f : \mathscr{X} \to \mathscr{R}$, where $\mathscr{X}$ is an appropriately valued domain. In our setting, we will take $\mathscr{X} = R^p$, but one could use more general spaces. A reproducing kernel K is a mapping $K : \mathscr{X} \times \mathscr{X} \to \mathscr{R}$ that satisfies the following properties: (a) $K(\cdot,x) \in \mathscr{F}$ for any $x \in \mathscr{X}$; (b) For any $f \in \mathscr{F}$ and any $x \in \mathscr{X}$, $f(x) =< f, K(\cdot,x) >_{\mathscr{F}}$. Property (b) is commonly referred to as the Reproducing property. There are several well-established equivalences regarding RKHS from the literature. We state them as a proposition here.
Proposition: The following are equivalent:

1. For any $x \in \mathscr{X}$, the function $\delta_x : \mathscr{F} \to \mathscr{R}$ defined by $\delta_x(f) = f(x)$ is continuous.
2. F has a reproducing kernel.
3. K is a kernel of a RKHS.

Next, we define a bivariate symmetric function $k(x,y)$ on $\mathscr{X} \times \mathscr{X}$ a kernel function if

$$\int_{\mathscr{X}} \int_{\mathscr{X}} k(x,y)g(x)g(y)dxdy \geq 0, \tag{5}$$

for all squared integrable functions $g(\cdot)$ on $\mathscr{X}$, i.e., $g(\cdot) \in L^2(\mathscr{X})$. By Mercer's theorem, existence of an RKHS is equivalent to the kernel function being positive definite. A positive definite kernel function is such that for any $Z_1, \ldots, Z_n \in \mathscr{X}$ and $c_1, \ldots, c_n \in R$,

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

Equivalently, a positive definite kernel induces an $n \times n$ positive definite matrix with (i,j)th entry $k(x_i, x_j)$.

Now suppose that $E[\sqrt{k(Z,Z)}] < \infty$. Then one can view the kernel function as being equivalent to a Hilbert-Schmidt operator (Bump, 1997, p. 592) so that one can then show that for a bounded and measurable kernel function,

$$\gamma_k(P_1, P_0) = \| \int k(\cdot, x) dP_1(x) - \int k(\cdot, x) dP_0(x) \|_{\mathscr{H}} = \|(P_1 - P_0)k\|$$

In words, this states that one can view γ_k as a pseudometric based on Hilbert space embeddings of P_0 and P_1.

A key property that is needed to achieve balance in covariates with kernels is that of being a characteristic kernel. This was proposed in Sriperumbudur, Fukumizu, Gretton, Schölkopf, Lanckriet et al. (2012) and means that $\gamma_k(P_1, P_0) = 0$ if and only if $P_1 = P_0$ for any pair of measures P_0, P_1 in a family of measures $\mathscr{P}$. Sriperumbudur et al. (2012) provide a simple characterization for characteristic kernels. Namely, integrally strictly positive definite kernel functions are sufficient to guarantee a kernel being characteristic. This involves replacing the inequality in (5) by strict inequality.

3 Proposed balance metric and computation

3.1 Kernel Distance

Zolotarev (1983) presents a hierarchy of probability metrics that have been used in the literature. The class of metrics we will work with is given by

$$\gamma(P,Q) = \sup_{f \in \mathscr{F}} | \int f dP - \int f dQ |, \tag{6}$$

where $\mathscr{F}$ is a class of functions. In (6), $\gamma(P,Q)$ is referred to by Zolotarev (1983) as an example of a probability metric with a $\zeta-$structure. We now give some examples for $\mathscr{F}$:

1. Let $\mathscr{F} = \{I_{(-\infty,t)} : t \in R^p\}$. Then (6) is the Kolmogorov distance.
2. Let $\|f\|_\infty = \sup_{x\in R^p} |f(x)|$. Then if $\mathscr{F} = \{f : \|f\|_\infty \leq 1\}$, (6) yields the total variation distance.
3. Define $\|f\|_L$ as
$$\|f\|_L = \sup\left\{ \frac{|f(x)-f(y)|}{\rho(x,y)} : x \neq y \in R^p \right\},$$
where ρ is a metric for R^p. $\|f\|_L$ is a Lipschitz metric, and setting $\mathscr{F} = \{f : \|f\|_L \leq 1\}$ in (6), this yields the Kantorovich metric. We note that a generalization of the Kantorovich metric is given by Fortet-Mourier metric, where $\|f\|_L$ is replaced by $\|f\|_C$, where
$$\|f\|_C = \sup\left\{ \frac{|f(x)-f(y)|}{c(x,y)} : x \neq y \in R^p \right\},$$
with $c(x,y) = \rho(x,y)\max(1, \rho(x,a)^{p-1}, \rho(y,a)^{p-1})$, where $p \geq 1$, $a \in R^p$.
4. Let $\|f\|_{BL} = \|f\|_\infty + \|f\|_L$. This is referred to as the Dudley metric, and setting $\mathscr{F} = \{f : \|f\|_{BL} \leq 1\}$, (6) yields the dual-bounded Lipschitz distance.
5. Let $\mathscr{K}$ denote an RKHS and define $\| \ \|_{\mathscr{K}}$ to be the norm for this space. Then setting $\mathscr{F} = \{f : \|f\|_{\mathscr{K}} \leq 1\}$ into (6), we get a 'kernelized' version of the total variation distance.

While many other choices of function classes are possible, a major outstanding issue is the feasibilty of computation of these metrics. This is related to the topic of multivariate goodness of fit statistics, whose computations become prohibitive in higher dimensions. However, it turns out that for RKHS, under a mild condition on the kernel, (6) has a closed-form solution. Let us define $T^* = {n_1}^{-1}$ if $T = 1$ and $T^* = -{n_0}^{-1}$ if $T = 0$, where n_1 is the sample size in the treatment group and n_0 is the sample size in the control group. We then have the following result, which is Theorem 2.4 of Sriperumbudur et al. (2012):

Theorem 1: Let k denote a strictly positive definite kernel function corresponding to the RKHS $\mathscr{K}$. Then
$$\gamma_k(P_{n1}, Q_{n0}) = \|\sum_{i=1}^{n} T_i^* k(\cdot, Z_i)\|_{\mathscr{K}} = \sqrt{\sum_{i,j=1}^{n} T_i^* T_j^* k(Z_i, Z_j)}, \tag{7}$$
where P_{n1} denote the empirical measure of $Z|T = 1$ and Q_{n0} denotes the empirical measure of $Z|T = 0$.

Equation (7) reveals a very simple, closed-form analytical solution for the empirical estimate of the probability metric under the assumption that the function class represents an RKHS. As pointed out in Srirempubudur et al., one can express the theoretical supremum in (6) as a linear combination of kernel functions under the same assumption in Theorem 1. Our proposal is to use (7) as a diagnostic for balance; smaller values of (7) denote better covariate balance. We also note from equation (7) that the computation of this balance statistic is $O(n^2)$ but it is also independent of the dimension of the confounders. This feature makes in appealing for applications in which there is a large number of covariates for which one needs to adjust for.

3.2 Choice of Kernel

An outstanding issue is the choice of RKHS for use in (7), or equivalently, the choice of $\mathscr{F}$. Clasically, in statistics, function spaces $\mathscr{F}$ have been chosen to be Sobolev spaces (Adams, 1975). One notable example of $\mathscr{F}$ is the Sobolev space corresponding to one-dimensional smoothing splines, which is given by

$$\mathscr{F}_{ss} = \{f : [a,b] \to R | \int_a^b \{f''(x)\}^2 < \infty\}$$

where f'' denotes the second derivative of the function f, and $[a,b]$ is a closed and finite interval in R. As is seen in the definition of $\mathscr{F}_{ss}$, Sobolev spaces are function spaces with constraints placed on either the function and/or its derivatives. Intuitively, the more constraints that are placed on the functions, the more restrictive the function class becomes. Put another way, with more derivative constraints that are placed, the fewer "directions" we search in computing the supremum statistic (7). Conversely, the less derivative restrictions that are in place, the bigger the function class will be. In this article, we will explore use a choice of K which comes from machine learning (Rasmussen and Williams, 2006) and is referred to as the Gaussian kernel. It is given by

$$K(x,y;\sigma^2) = \exp(-\|x-y\|^2/\sigma^2),$$

where $\sigma^2 > 0$ is a parameter to either be fixed or estimated from the data. In this article, we will use the median or mean of all possible pairwise squared Euclidean distances between all pairs of subjects to estimate σ^2. It is well-known that the Gaussian kernel corresponds to a Gaussian stochastic process with infinitely differentiable paths. Steinwart et al. (2006), Theorem 3.4, gives a characterization the function space corresponding to the this kernel.

Theorem 2: The function space corresponding to the Gaussian kernel is given by

$$\mathscr{F}_G = \{f : R^p \to R| \sum_{m=0}^{\infty} \frac{\sigma^{2m}}{m!2m} (D^m f)^2 < \infty\},$$

where $D^{2m} = \bigtriangledown^{2m} f$, $D^{2m+1} f = \triangle(\bigtriangledown^{2m} f)$, $\triangle$ is the gradient operator and $\bigtriangledown^{2m}$ is the Laplacian operator applied m times.

Thus, we see that the Gaussian kernel puts constraints on all orders of derivatives so that there will be fewer functions in this function space. However, the kernel distance based on the Gaussian kernel will seek to simultaneously satisfy all derivative constraints and thus will have a strict definition for balance.

4 Simulation studies

In this section, we conduct simulation studies to investigate the performance of the balance metric based on kernels and compare it with other commonly used balance statistics in the causal inference literature. We follow the simulation study by Austin, Grootendorst, and Anderson (2007), Belitser, Martens, Pestman, Groenwold, Boer, and Klungel (2011) and Stuart, Lee, and Leacy (2013). First, with a sample size of $n = 1000$, we generate nine covariates, Z_1, Z_3, Z_4, Z_5, Z_8 from $N(0,1)$ and Z_2, Z_6, Z_7, Z_9 from Bernoulli(0.5). Then, the treatment variable T is generated from Bernoulli$(e(Z))$ where

$$\begin{aligned} \text{logit}(e(Z)) =& \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_4 + \alpha_4 Z_5 + \alpha_5 Z_7 + \alpha_6 Z_8 + \alpha_7 Z_2 Z_4 \\ &+ \alpha_8 Z_2 Z_7 + \alpha_9 Z_7 Z_8 + \alpha_{10} Z_4 Z_5 + \alpha_{11} Z_1 Z_1 + \alpha_{12} Z_7 Z_7, \end{aligned}$$

and

$$\begin{aligned} \alpha =& (0, \log(2), \log(1.4), \log(2), \log(1.4), \log(2), \log(1.4), \log(1.2), \log(1.4), \\ & \log(1.6), \log(1.2), \log(1.4), \log(1.6)). \end{aligned}$$

The outcome variable Y is generated from four different scenarios which differ in terms of model complexity:

$$\begin{aligned} A : Y =& \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 Z_6 + \gamma T; \\ B : Y =& \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_7 Z_2 Z_4 + \beta_8 Z_3 Z_5 \\ &+ \beta_9 Z_3 Z_6 + \beta_{10} Z_4 Z_5 + \gamma T; \\ C : Y =& \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_{11} Z_1 Z_1 + \beta_{12} Z_6 Z_6 + \gamma T; \\ D : Y =& \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_7 Z_2 Z_4 + \beta_8 Z_3 Z_5 \\ &+ \beta_9 Z_3 Z_6 + \beta_{10} Z_4 Z_5 + \beta_{11} Z_1 Z_1 + \beta_{12} Z_6 Z_6 + \gamma T, \end{aligned}$$

where $\beta = (-2.4, 1.68, 1.68, 1.68, 3.47, 3.47, 3.47, 0.91, 1.68, 2.35, 0.91, 1.68, 2.35)$ and the true causal effect is a constant: $\gamma = 3$.

In each simulation, we fit forty different propensity score models which are exactly the same as in the simulation study in Belitser et al. (2011). We then employ one-to-one matching based on the estimated propensity scores to estimate the average causal effect among the treated (ATT). We calculate the absolute bias of the estimators based on different propensity score models and the corresponding balance metric values based on the matched data. We then calculate the Pearson correlation coefficient between the absolute bias and the balance metric. We use equation (7) to calculate the kernel distance and let k to be the Gaussian kernel. To compare with the proposed balance metric, we also calculate the absolute standardized mean difference (ASMD) of a given covariate and then look at the mean, maximum and median of the ASMD values over all nine covariates. The other balance matrics we are going to compare are the average Kolmogorov-Smirnov (KS) test statistic and the average t-test statistic. To be noticed, Stuart et al. (2013) proposed a new balance measure based on the prognostic score, which involves the modeling of a outcome model in the control group and shows superior performance in simulation studies. However, in our study, we are going to focus only on model-free balance metrics. We repeat the process 1000 times and report the average correlation coefficients and the standard deviation of the correlation coefficients. The results are displayed in the first part of Table 1.

In the second set of simulations, we use the same model setup as in the first simulation; the only difference is that instead of performing matching, we employ inverse probability weighting (IPW) to estimate ATT. In this case, the proposed balance metric is still calculated as (7), but we need to redefine T_i^*:

$$T_i^* = \frac{w_i}{\sum_{j=1}^n T_j w_j}$$

if $T_i = 1$, and

$$T_i^* = -\frac{w_i}{\sum_{j=1}^n (1 - T_j) w_j}$$

if $T_i = 0$, where w_i is the inverse probability weight for subject i. Since we focus on ATT, $w_i = 1$ if $T_i = 1$ and $w_i = \hat{e}(Z_i)/(1 - \hat{e}(Z_i))$ if $T_i = 0$, where $\hat{e}(Z)$ is the estimated propensity score from a particular propensity score model. The simulation results are displayed in the second part of Table 1.

In the third set of simulations, we perform subclassification to estimate the causal effect instead of matching and IPW. We divide the sample into five strata based on the quantiles of the estimated propensity scores and estimate ATT in each stratum. The final causal estimate is the average value of the five estimates. The

Table 1: Mean and standard deviation of the Pearson correlation coefficients

	Matching: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.632 (0.134)	0.609 (0.155)	0.606 (0.152)	0.587 (0.156)
max ASMD	0.557 (0.176)	0.542 (0.196)	0.548 (0.185)	0.512 (0.193)
median ASMD	0.367 (0.214)	0.351 (0.212)	0.347 (0.221)	0.356 (0.213)
mean KS	0.372 (0.264)	0.358 (0.267)	0.348 (0.276)	0.333 (0.273)
mean t-statistic	0.634 (0.133)	0.609 (0.155)	0.610 (0.149)	0.588 (0.155)
kernel distance	0.797 (0.115)	0.773 (0.140)	0.788 (0.120)	0.759 (0.125)
	Inverse Probability Weighting: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.717 (0.092)	0.668 (0.125)	0.693 (0.100)	0.664 (0.136)
max ASMD	0.642 (0.136)	0.611 (0.140)	0.607 (0.155)	0.592 (0.180)
median ASMD	0.469 (0.191)	0.441 (0.195)	0.462 (0.173)	0.459 (0.204)
mean KS	0.595 (0.163)	0.539 (0.183)	0.554 (0.151)	0.519 (0.214)
mean t-statistic	0.716 (0.093)	0.668 (0.125)	0.694 (0.100)	0.661 (0.133)
kernel distance	0.839 (0.085)	0.808 (0.100)	0.808 (0.098)	0.780 (0.137)
	Subclassification: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.655 (0.114)	0.649 (0.119)	0.630 (0.123)	0.617 (0.127)
max ASMD	0.613 (0.150)	0.611 (0.154)	0.594 (0.158)	0.578 (0.161)
median ASMD	0.527 (0.180)	0.521 (0.181)	0.506 (0.184)	0.500 (0.186)
mean KS	0.455 (0.194)	0.450 (0.197)	0.427 (0.199)	0.414 (0.203)
mean t-statistic	0.677 (0.103)	0.670 (0.109)	0.655 (0.110)	0.639 (0.114)
kernel distance	0.833 (0.082)	0.827 (0.086)	0.816 (0.091)	0.798 (0.092)

mean and the standard deviation of the correlation coefficients between the absolute bias and the balance measures are displayed in the third part of Table 1.

Harder et al. (2010) stated that by achieving balance, the bias in the estimated causal treatment effect due to measured covariates can be reduced. Table 1 shows that the balance metric based on kernel distance has the largest correlation with the absolute bias in the estimated casual effect, which means this metric is the best indicator of bias in the estimation of causal effects. In practice, the true causal effect is unknown and to evaluate the goodness of a matching, IPW or subclassification procedure, one can check the balance in the covariates and the simulation study indicates one of the best criteria is to use the proposed kernel distance. In addition, the kernel distance also has the smallest standard deviation.

We next check whether the propensity score model picked by the proposed balance metric can actually lead to reduced bias in estimating ATT, compared to existing balance metrics. We follow the same simulation setup as in the previous subsection. In each method, we employ a particular balance metric to select the optimal model and perform IPW/matching/sublclassification using the chosen propensity score model. We report the bias and the standard deviation of the estimated causal effect in Table 2, which shows the proposed balance metric can lead to the least biased and variable estimation of the causal effects.

5 Data Application

In this section, we are going to apply the proposed methodology to a longitudinal study: the Early Dieting in Girls Study (Fisher and Birch, 2002). There are multiple factors that may influence children's eating behavior, of which mother's eating behavior and attitudes is one of the most important factors (Sinton and Birch, 2005). It has been shown in the literature that mothers may influence daughters' dieting by endorsing dieting themselves and by providing information and encouragement about dieting (e.g., Benedikt, Wertheim, and Love, 1998, Birch and Fisher, 2000, Neumark-Sztainer, Bauer, Friend, Hannan, Story, and Berge, 2010). However, these studies are correlational studies and none of them aim to draw causal inference between mother's weight concern and girl's dieting behavior. The motivating question in this analysis is whether mother's overall weight concern increases the likelihood of early dieting behavior among girls at Age 7. The participants in this study are 197 daughters and their mothers, who are from non-Hispanic, White families living in central Pennsylvania. Both daughters and their mothers were interviewed at daughters' age five (Wave 1), seven (Wave 2), nine (Wave 3), eleven (Wave 4), thirteen (Wave 5) and fifteen (Wave 6). At each wave, they paid a scheduled visit to the laboratory and filled questionnaires.

Table 2: Bias and standard deviation of the estimated ATT

	Matching: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.303 (0.783)	0.395 (1.134)	0.983 (0.603)	1.154 (0.948)
max ASMD	0.291 (0.799)	0.357 (1.143)	0.960 (0.617)	1.136 (1.004)
median ASMD	1.556 (1.819)	1.865 (2.094)	1.990 (1.539)	2.539 (2.098)
mean KS	0.719 (1.318)	0.851 (1.608)	1.272 (1.070)	1.507 (1.419)
mean t-statistic	0.294 (0.783)	0.383 (1.130)	0.970 (0.594)	1.141 (0.948)
kernel distance	0.239 (0.754)	0.356 (1.061)	0.924 (0.576)	1.059 (0.904)

	Inverse Probability Weighting: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.622 (0.637)	0.751 (0.858)	1.283 (0.526)	1.555 (0.786)
max ASMD	0.637 (0.699)	0.788 (0.946)	1.273 (0.555)	1.589 (0.858)
median ASMD	1.756 (1.545)	2.032 (1.791)	2.211 (1.307)	2.625 (1.666)
mean KS	0.759 (0.699)	0.930 (0.926)	1.374 (0.566)	1.711 (0.818)
mean t-statistic	0.627 (0.639)	0.760 (0.862)	1.283 (0.526)	1.562 (0.790)
kernel distance	0.596 (0.598)	0.733 (0.805)	1.283 (0.502)	1.549 (0.732)

	Subclassification: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.689 (0.718)	0.819 (0.950)	1.501 (0.711)	1.613 (0.992)
max ASMD	0.679 (0.700)	0.768 (0.945)	1.480 (0.723)	1.596 (0.987)
median ASMD	1.105 (1.110)	1.264 (1.367)	1.866 (1.059)	2.120 (1.438)
mean KS	0.842 (0.881)	0.978 (1.145)	1.614 (0.842)	1.722 (1.114)
mean t-statistic	0.687 (0.700)	0.823 (0.950)	1.507 (0.703)	1.594 (0.978)
kernel distance	0.590 (0.428)	0.676 (0.625)	1.220 (0.364)	1.432 (0.519)

The treatment variable in this analysis is mother's overall weight concern, which is calculated as the average score of five Likert scale questions. It is a summary of mother's concern about gaining weight before Wave 2. A higher value implies the mother is more concerned about gaining weight. In the dataset, its values range from 0 to 3.4. Since we focus on binary treatments, we first dichotomize the variable at its median in the sample, which has a value of 1.6. After dichotomizing the variable, a value of 1 implies the mother has high weight concern and a value of 0 implies the mother has low weight concern. This interpretation also makes sense in the context of the questions in the questionnaire. For example, one of the Likert scale questions is "How afraid are you of gaining 3 pounds?" A chosen value that is greater than 1.6 means the mother is at least moderately afraid of gaining 3 pounds while a chosen value less or equal to 1.6 means the mother is not afraid or slightly afraid of that. The outcome variable is *earlydiet*, which is the indicator of girl's early dieting behavior at age 7. There are 49 baseline covariates in the study, which are measured at Wave 1 (girls' age 5). We first fit a univariate logistic regression model of the treatment/outcome variable on each covariate. Based on the Wald test at $\alpha = 0.05$, 22 variables are not related to either the treatment or the outcome variable. The rest of the covariates are divided into three groups as shown in Table 3, depending on whether the variable is significantly related to the treatment or the outcome variable. To be noticed, the covariates in set 3 are real confounders in the sense they are significantly related to both the treatment and the outcome variable.

To draw causal inference, we perform a one-to-one matching with replacement using genetic matching (Diamond and Sekhon, 2013). In genetic matching, the generalized Mahalanobis distance (GMD) between subject i and subject j is calculated as

$$GMD(Z_i, Z_j, W) = \sqrt{(Z_i - Z_j)'(S^{-1/2})'WS^{-1/2}(Z_i - Z_j)}, \tag{8}$$

where S is sample covariance matrix of Z. W is the diagnoal weight matrix where the ith diagnoal element is the weight placed on the ith covariate while measuring the distance. The algorithm iteratively updates W, i.e., the weight for each covariate, while performing multivariate matching, until a certain balance metric based on the matched dataset is minimized. For this dataset, we compare genetic matching with the objective to minimize the kernel distance to other existing balance measurements. The matching procedure is implemented by the *Matching* package in R (Sekhon, 2011). In Table 4, "qqmean.mean" refers to genetic matching by minimizing the mean standardized difference in the empirical QQ plot for each covariate. Similarly, "qqmedian.median" focuses on minimizing the median standardized difference while "qqmax.max" focuses on minimizing the maximum standardized difference. "pvals" focuses on maximizing the p-values from t-test and KS test.

Table 3: List of variables in the Early Dieting in Girls study

Variable Set	Description	Variables
Set 1	Only related to T	mother's baseline depression, self-esteem, satisfaction of current body, perceived competence in eating scale (effort, importance), perceived overweight subscale score, disinhibition score, hunger score, restraint score; girl's baseline disinhibition score.
Set 2	Only related to Y	mother's baseline satisfaction with girl's body, mother's opinion of body characteristic inheritance, report of girl's overall pubertal development, restriction subscale score; girl's baseline BMI Z-score, overall total body esteem, overweight based on NCHS 2000, overall weight-related teasing
Set 3	Related to T and Y (real confounders)	mother's baseline BMI, currently dieting to lose weight, overall weight, perception of current size, external locus of control, role overload, total weight-related teasing, weight concerns subscale score.

As discussed in Diamond and Sekhon (2013), the covariates in Set 1 are instruments in the sense they are not significantly related to the outcome variable but highly predictive of the treatment. Matching on instruments may increase the bias and variance of the causal effect estimator (Stuart, 2010, Zhu, Schonbach, Coffman, and Williams, 2015b). Therefore, the covariates we aim to balance are those in Set 2 and Set 3 listed in Table 3. We use different balance criteria to obtain the optimal matched datasets and then regress the outcome variable on the treatment variable using the matched dataset. The causal log odds ratio estimates are displayed in Table 4. The confidence intervals and the standard errors for the causal log odds ratio are reported on 100 bootstrapped samples. Among all estimators based on genetic matching, the estimator with the objective to minimize kernel distance yields the smallest standard error. The analyses suggest that the mother's weight concern increases the child's probability of early dieting behavior. However, because all the matching-based estimators have 95% confidence intervals that contain zero, the causal effect is not statistically significant at a 0.05 significance level. This conclusion is consistent with the findings in Zhu, Coffman, and Ghosh (2015a), where mother's overall weight concern is treated as a continuous treatment variable.

Table 4: Causal Estimates for Different Approaches

Method	Estimate	SE	95% CI	kernel distance
Genetic matching				
kernel distance	0.221	0.454	(-0.522, 1.100)	0.060
qqmean.mean	0.137	0.461	(-0.761, 1.093)	0.064
qqmedian.median	0.153	0.508	(-0.781, 1.054)	0.073
qqmax.max	0.179	0.493	(-0.688, 1.140)	0.065
pvals	0.144	0.468	(-0.677, 1.059)	0.067
Without matching	0.987	0.421	(0.365, 1.833)	0.250

To evaluate the goodness of the matching procedures, we also report the mean ASMD and the kernel distance for each matched sample. As shown in Table 4, the genetic matching procedure by minimizing the kernel distance yields the smallest kernel distance value. We also plot the ASMD value for each covariate in Figure 1 As shown in the figure, only for kernel distance, all the covariates are balanced with ASMD values smaller than 0.2, the cut-off value.

Based on Gretton et al. (2012, Corollary 9), we can suggest a cut-off value for the kernel distance, which is $\sqrt{2K/n}(1+\sqrt{2\log\alpha^{-1}})$, where n is the sample size in the treatment/control group for the one-to-one matched dataset and α is a pre-specified significance level, and K is the upper bound for the kernel function,

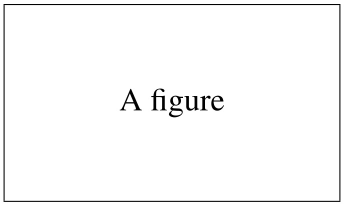

Figure 1: ASMD values before and after matching for different balance metrics in genetic matching

which equals 1 if we employ the Gaussian kernel. This formula is used to determine the acceptance region for a kernel two-sample test based on equation (7). Since we are not conducting hypothesis testing here, we prefer a large value of α. Based on different significance levels (α=0.05, 0.2, 0.5), the cut-off value is 0.3474, 0.2815 and 0.2194, respectively. The kernel distance for the best matched dataset is 0.060, which is below the cut-off values at different significance levels. Therefore, we conclude the data achieves overall balance in the covariates.

6 Discussion

In this article, we have developed a simple statistic for diagnosing covariate balance in observational studies. The methodology is based on reproducing kernel Hilbert space theory and shows one of the applications of this machine learning approach to an important problem in statistics: checking balance. Using a standard kernel from the machine learning literature, the Gaussian kernel, the proposed balance metric is shown to outperform the existing commonly used balance statistics. The DCB condition allows for multivariate comparison of the joint distribution of confounders. This condition, coupled with use of the RKHS approach, leads to a computationally tractable approach to analysis that allows for relatively large numbers of concomitant variables.

There are several directions that are worthy of further investigation. One would be to see if one can use (6) in manner similar to what was done in Imai and Ratkovic (2014) in order to create a propensity score estimation procedure that will satisfy DCB. If such a method were possible, then this would lead to propensity score models which would be ideally tailored for performing causal inference.

Acknowledgments

The project described was supported by Award Number 430-2016-00163 from the Social Sciences and Humanities Research Council. The content is solely the responsibility of the authors.

References

Adams, R. (1975): *Sobolev Spaces*, Academic Press, New York.

Austin, P., P. Grootendorst, and G. Anderson (2007): "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study," *Statistics in medicine*, 26, 734–753.

Austin, P. C. and E. A. Stuart (2015): "Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies," *Statistics in medicine*, 34, 3661–3679.

Belitser, S., E. Martens, W. Pestman, R. Groenwold, A. Boer, and O. Klungel (2011): "Measuring balance and model selection in propensity score methods," *Pharmacoepidemiology and drug safety*, 20, 1115–1129.

Benedikt, R., E. Wertheim, and A. Love (1998): "Eating attitudes and weight-loss attempts in female adolescents and their mothers," *Journal of Youth and Adolescence*, 27, 43–57.

Berlinet, A. and C. Thomas-Agnan (2011): *Reproducing kernel Hilbert spaces in probability and statistics*, Springer Science & Business Media.

Birch, L. and J. Fisher (2000): "Mothers' child-feeding practices influence daughters' eating and weight," *The American Journal of Clinical Nutrition*, 71, 1054–1061.

Bump, D. (1997): "Automorphic forms and representations," .

Diamond, A. and J. Sekhon (2013): "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95, 932–945.

Fisher, J. O. and L. Birch (2002): "Eating in the absence of hunger and overweight in girls from 5 to 7 y of age," *The American journal of clinical nutrition*, 76, 226–231.

Hainmueller, J. (2011): "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, mpr025.

Harder, V., E. Stuart, and J. Anthony (2010): "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research." *Psychological Methods*, 15, 234–249.

Hazlett, C. (2015): "Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects," *Available at SSRN 2746753*.

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

Holland, P. (1986): "Statistics and causal inference," *Journal of the American statistical Association*, 81, 945–960.

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

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

Neumark-Sztainer, D., K. Bauer, S. Friend, P. Hannan, M. Story, and J. Berge (2010): "Family weight talk and dieting: how much do they matter for body dissatisfaction and disordered eating behaviors in adolescent girls?" *Journal of Adolescent Health*, 47, 270–276.

Neyman, J. (1923): "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51.

Rachev, S., L. Klebanov, S. Stoyanov, and F. Fabozzi (2013): *The methods of distances in the theory of probability and statistics*, Springer Science & Business Media.

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

Rubin, D. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688–701.

Sekhon, J. (2011): "Multivariate and propensity score matching software with automated balance optimization: the matching package for r," *Journal of Statistical Software*, 42.

Sinton, M. and L. Birch (2005): "Weight status and psychosocial factors predict the emergence of dieting in preadolescent girls," *International Journal of Eating Disorders*, 38, 346–354.

Sriperumbudur, B., K. Fukumizu, A. Gretton, B. Schölkopf, G. Lanckriet, et al. (2012): "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, 6, 1550–1599.

Steinwart, I., D. Hush, and C. Scovel (2006): "An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels," *Information Theory, IEEE Transactions on*, 52, 4635–4643.

Stuart, E., B. Lee, and F. Leacy (2013): "Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research," *Journal of Clinical Epidemiology*, 66, S84–S90.

Stuart, E. A. (2010): "Matching methods for causal inference: A review and a look forward," *Statistical Science*, 25, 1–21.

Wahba, G. (1990): *Spline models for observational data*, volume 59, Siam.

Zhu, Y., D. Coffman, and D. Ghosh (2015a): "A boosting algorithm for estimating generalized propensity scores with continuous treatments," *Journal of Causal Inference*, 3, 25–40.

Zhu, Y., M. Schonbach, D. L. Coffman, and J. S. Williams (2015b): "Variable selection for propensity score estimation via balancing covariates," *Epidemiology*, 26, e14–e15.

Zolotarev, V. (1983): "Probability metrics," *Theory Probab. Appl.*, 28, 264–287.