Yeying Zhu*, Jennifer S. Savage, and Debashis Ghosh

A Kernel-Based Metric for Balance Assessment

https://doi.org/10.1515/jci-2016-0029
Received December 18, 2016; revised April 2, 2018; accepted May 2, 2018

Abstract: An important goal in causal inference is to achieve balance in the covariates among the treatment groups. In this article, we introduce the concept of distributional balance preserving which requires the distribution of the covariates to be the same in different treatment groups. We also introduce a new balance measure called kernel distance, which is the empirical estimate of the probability metric defined in the reproducing kernel Hilbert spaces. Compared to the traditional balance metrics, the kernel distance measures the difference in the two multivariate distributions instead of the difference in the finite moments of the distributions. Simulation results show that the kernel distance is the best indicator of bias in the estimated casual effect compared to several commonly used balance measures. We then incorporate kernel distance into genetic matching, the state-of-the-art matching procedure and apply the proposed approach to analyze the Early Dieting in Girls study. The study indicates that mothers' overall weight concern increases the likelihood of daughters' early dieting behavior, but the causal effect is not significant.

Keywords: Causal effect, Distributional covariate balance, Probability metric, Reproducing kernel Hilbert space

1 Introduction

Determining causal effects from observational studies continues to be of great interest to scientists and doctors. 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 [1, 2]. Within this setup, an important quantity to facilitate causal inference has been the propensity score [3], 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 [3], 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 [4], Sekhon [5] and Iacus, King, and Porro [6]. Belitser, Martens, Pestman, Groenwold, Boer, and Klungel [7] proposed using

***Corresponding author: Yeying Zhu,** University of Waterloo, Department of Statistics and Actuarial Science, 200 University Ave W, Waterloo, Ontario, N2L 3G1, Canada, e-mail: yeying.zhu@uwaterloo.ca
Jennifer S. Savage, Pennsylvania State University, Center for Childhood Obesity Research, University Park, Pennsylvania, United States, e-mail: jfs195@psu.edu
Debashis Ghosh, University of Colorado School of Public Health, Biostatistics and Informatics, 13001 E. 17th Place, Aurora, 80045, Colorado, United States, e-mail: debashis.ghosh@ucdenver.edu

the overlapping coefficient, the Kolmogorov-Smirnov (KS) distance or the Lévy distance to measure balance. Franklin, Rassen, Ackermann, Bartels, and Schneeweiss [8] compared several balance metrics through simulation studies based on the matching procedure. Austin and Stuart [9] 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 [10]. Examples include the GBM [11], entropy balancing [12], CBPS [13], kernel balancing [14] and the model-averaging approach [15]. In some of these approaches, a balance statistic, such as the average standardized absolute mean difference, c-statistic or KS 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 observational studies, it is well-acknowledged that the difference in the observed outcomes between the treatment group and the control group is generally biased for estimating the causal treatment effects. The bias depends on the relationship between the potential outcomes and the covariates, as well as the distribution of the covariates in different treatment groups [16]. In Hazlett [14], 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 is not enough. This motivates us to propose a balance metric based on the multivariate distribution of the covariates instead of the finite moments of the covariates.

The layout of this paper is as follows. 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 kernel distance against several other balance metrics 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 [2, 17] 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 [3]:

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

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 [3] 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 [3] 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 $Z|T = 0$. We now review the concept of probability metrics, an overview for which can be found in Zolotarev [18]:

Definition 1. Let X, V, W 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 $P(X = V) = 1$, then $d(X, V) = 0$.
(PM2) $d(X, V) = d(V, X)$.
(PM3) $d(X, W) \leq d(X, V) + d(V, W)$.

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 [18] and Rachev, Klebanov, Stoyanov, and Fabozzi [19] give comprehensive overviews on probability metrics. Given the definition of probability metric, we can now define distributional covariate balance (DCB) as $d(P, Q) = 0$ where P and Q 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 [13], 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 [20], Berlinet and Thomas-Agnan [21] and Steinwart, Hush, and Scovel [22]. 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.

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 $x_1, \ldots, x_n \in \mathscr{X}$ 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 $k(x_i, x_j)$.

In the next section, we are going to define a particular probability metric and give several examples of the metric based on different functional classes. We then focus on one particular form: kernel distance based on RKHS.

3 Proposed balance metric and its computation

3.1 Kernel distance

Zolotarev [18] 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 [18] as an example of a probability metric with a ζ-structure. We now give some examples for $\mathscr{F}$, which are also reviewed in Sriperumbudur, Fukumizu, Gretton, Schölkopf, Lanckriet et al. [23]:

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 feasibility 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 empirical estimator. Therefore, we are going to focus on the probability metric based on RKHS as shown in the fifth example above. We call it kernel distance and denote it as $\gamma_k(P, Q)$.

Now suppose that $E[\sqrt{k(X, X)}] < \infty$. Then one can view the kernel function as being equivalent to a Hilbert-Schmidt operator ([24], p. 592) so that one can then show that for a bounded and measurable kernel function,

$$\gamma_k(P, Q) = \| \int k(\cdot, x) dP(x) - \int k(\cdot, x) dQ(x) \|_{\mathscr{H}}$$

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

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, Gretton, Fukumizu, Schölkopf, and Lanckriet [25] and means that $\gamma_k(P, Q) = 0$ if and only if $P = Q$ for any pair of measures P, Q. Sriperumbudur et al. [25] 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.

Now we are ready to introduce the empirical estimator of $\gamma_k(P, Q)$. 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. [23]:

Theorem 1. *Let k denote a strictly positive definite kernel function corresponding to the RKHS $\mathscr{K}$. Then an empirical estimator of (6) is*

$$\gamma_k(P_{n1}, Q_{n0}) = \| \sum_{i=1}^{n} T_i^* k(\cdot, Z_i) \|_{\mathscr{K}} = \sqrt{\sum_{i,j=1}^{n} T_i^* T_j^* k(Z_i, Z_j)}, \tag{7}$$

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 Sriperumbudur et al. [23], 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 $\mathscr{F}$. Classically, in statistics, function spaces $\mathscr{F}$ have been chosen to be Sobolev spaces [26]. 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 (6). Conversely, the less derivative restrictions that are in place, the bigger the function class will be. In this article, we will use a choice of k which comes from machine learning 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. When all the covariates are standardized, an alternative is to use the dimension of the covariates to estimate σ^2, which is equivalent to the expected value of the pairwise Euclidean distance [14]. It is well-known that the Gaussian kernel corresponds to a Gaussian stochastic process with infinitely differentiable paths. Steinwart et al. [22], Theorem 3.4, gives a characterization of 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.

To further understand the kernel distance with k to be the Gaussian kernel, we consider a concrete example. Let us assume $Z|T = 1$ follows $N(\mu I_{d\times 1}, \sigma^2 I_{d\times d})$, i. e., $P = N(\mu I_{d\times 1}, \sigma^2 I_{d\times d})$ and $Z|T = 0$ follows $N(\lambda I_{d\times 1}, \theta^2 I_{d\times d})$, i. e., $Q = N(\lambda I_{d\times 1}, \theta^2 I_{d\times d})$, where d is the dimension of the covariates. Let $k(x, y) = \exp(-\|x - y\|^2/2d)$. As a special example of Sriperumbudur et al. [23], we find the probability metric (6) boils down to a simple form:

$$\gamma_k(P, Q) = \sqrt{\left(\frac{d}{2\sigma^2 + d}\right)^{d/2} + \left(\frac{d}{2\theta^2 + d}\right)^{d/2} - 2\left(\frac{d}{\sigma^2 + \theta^2 + d}\right)^{d/2} e^{-\frac{d(\mu-\lambda)^2}{2(\sigma^2+\theta^2+d)}}} \tag{8}$$

As can be seen from (8), the kernel distance incorporates the location parameters and the scale parameters into one number in a highly nonlinear fashion. We then generate $n_1 = 500$ data points from the treatment group and $n_0 = 500$ data points from the control group following the above distribution P and Q, respectively. Then, we use (7) to calculate the empirical estimator of the kernel distance. We generate 100 datasets and calculate the average value of $\gamma_k(P_{n1}, Q_{n0})$. The true and empirical estimates of kernel distance under different combinations of the values of the parameters are displayed in Table 1. It is shown that the empirical estimator can approximate the true kernel distance for different values of d, which guarantees that we can use (7) to measure discrepancy/balance between two distributions even under moderate dimensions.

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 [27], Belitser et al. [7] and Stuart, Lee, and

Table 1: True and Empirical Estimates of Kernel Distance Under Various Settings.

Setting	$d = 1$		$d = 5$		$d = 10$		$d = 50$	
	True	Empirical	True	Empirical	True	Empirical	True	Empirical
$\mu = 0, \lambda = 1, \sigma = \theta = 1$	0.421	0.420	0.509	0.485	0.523	0.490	0.535	0.495
$\mu = 0, \lambda = 2, \sigma = \theta = 1$	0.748	0.750	0.810	0.798	0.807	0.652	0.800	0.806
$\mu = 0, \lambda = 3, \sigma = \theta = 1$	0.947	0.950	0.910	0.976	0.886	0.978	0.860	0.974
$\mu = 0, \lambda = 1, \sigma = 1, \theta = \sqrt{2}$	0.377	0.376	0.458	0.427	0.470	0.433	0.480	0.427
$\mu = 0, \lambda = 2, \sigma = 1, \theta = \sqrt{2}$	0.647	0.658	0.696	0.709	0.687	0.708	0.671	0.698
$\mu = 0, \lambda = 3, \sigma = 1, \theta = \sqrt{2}$	0.837	0.862	0.790	0.886	0.756	0.881	0.717	0.870
$\mu = 0, \lambda = 1, \sigma = 1, \theta = \sqrt{3}$	0.382	0.378	0.471	0.418	0.487	0.414	0.500	0.400
$\mu = 0, \lambda = 2, \sigma = 1, \theta = \sqrt{3}$	0.596	0.612	0.647	0.419	0.639	0.650	0.623	0.634
$\mu = 0, \lambda = 3, \sigma = 1, \theta = \sqrt{3}$	0.769	0.809	0.730	0.827	0.694	0.815	0.655	0.802

Leacy [28]. 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

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

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:

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

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: $\gamma = 3$. Here, Y is generated without an error term. The randomness comes from the treatment model and the generation of Z.

In each simulation, we fit forty different propensity score models which are exactly the same as in the simulation study in Belitser et al. [7]. We then employ one-to-one matching with replacement based on the estimated propensity scores to estimate the average causal effect among the treated (ATT). The matching procedure is done without calipers, i. e., a treated subject will be matched to the nearest control subject no matter how far the two subjects are. Once the matched pairs are found, the ATT estimate is simply the difference in the mean outcomes between the treatment group and the control group. 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 forty pairs of 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. We repeat the process 1000 times and report the average Pearson correlation coefficients and the standard deviation of the correlation coefficients.

Table 2: 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)	0.605 (0.152)	0.587 (0.155)
max ASMD	0.556 (0.176)	0.541 (0.198)	0.548 (0.185)	0.512 (0.192)
median ASMD	0.367 (0.213)	0.351 (0.213)	0.347 (0.220)	0.357 (0.212)
mean KS	0.372 (0.265)	0.359 (0.267)	0.347 (0.277)	0.333 (0.272)
mean t-statistic	0.633 (0.134)	0.609 (0.155)	0.609 (0.150)	0.589 (0.154)
kernel distance	0.801 (0.109)	0.777 (0.132)	0.789 (0.118)	0.759 (0.124)
prognostic score	0.985 (0.013)	0.955 (0.043)	0.961 (0.034)	0.954 (0.038)
	Inverse Probability Weighting: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.692 (0.096)	0.677 (0.111)	0.683 (0.112)	0.663 (0.121)
max ASMD	0.606 (0.158)	0.596 (0.174)	0.615 (0.160)	0.585 (0.167)
median ASMD	0.468 (0.185)	0.458 (0.187)	0.449 (0.194)	0.449 (0.190)
mean KS	0.564 (0.157)	0.553 (0.166)	0.539 (0.183)	0.521 (0.183)
mean t-statistic	0.694 (0.093)	0.678 (0.109)	0.685 (0.108)	0.664 (0.117)
kernel distance	0.822 (0.092)	0.809 (0.105)	0.812 (0.105)	0.785 (0.113)
prognostic score	0.994 (0.009)	0.981 (0.033)	0.979 (0.039)	0.977 (0.036)
	Subclassification: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.598 (0.171)	0.584 (0.180)	0.581 (0.183)	0.567 (0.188)
max ASMD	0.528 (0.207)	0.520 (0.219)	0.519 (0.215)	0.491 (0.220)
median ASMD	0.296 (0.245)	0.290 (0.247)	0.279 (0.248)	0.275 (0.247)
mean KS	0.182 (0.324)	0.176 (0.325)	0.165 (0.327)	0.162 (0.327)
mean t-statistic	0.640 (0.141)	0.623 (0.154)	0.628 (0.150)	0.606 (0.157)
kernel distance	0.769 (0.166)	0.763 (0.168)	0.747 (0.181)	0.730 (0.180)
prognostic score	0.940 (0.038)	0.905 (0.066)	0.924 (0.047)	0.898 (0.062)

The results are displayed in the first part of Table 2. In case the relationship between the bias and the balance metric is nonlinear, we also report the average Spearman's rank correlation coefficient and its standard deviation in Table 3. We find that the ranking of the relative performance of different methods are the same for either Pearson correlation or Spearman's rank correlation.

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 2 for Pearson correlation coefficient and in the second part of Table 3 for Spearman's rank correlation coefficient.

In the third set of simulations, we perform subclassification to estimate the causal effect instead of matching and IPW. We first divide the subjects in the treatment group into five strata based on the percentiles of the propensity scores. Based on the percentiles, we then find the control subjects in each stratum. If the estimated propensity score for a subject in the control group falls out of the range of the propensity scores in the treatment group, the subject will be discarded. In each stratum, we estimate ATT by calculating the difference in the mean outcome between the treatment group and the control group. The final causal estimate

Table 3: Mean and standard deviation of the Spearman's rank correlation coefficients.

	Matching: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.557 (0.149)	0.548 (0.161)	0.533 (0.172)	0.534 (0.168)
max ASMD	0.481 (0.225)	0.484 (0.231)	0.480 (0.239)	0.462 (0.239)
median ASMD	0.283 (0.233)	0.273 (0.232)	0.243 (0.244)	0.268 (0.234)
mean KS	0.319 (0.257)	0.315 (0.259)	0.293 (0.280)	0.296 (0.270)
mean t-statistic	0.558 (0.148)	0.549 (0.159)	0.538 (0.167)	0.536 (0.164)
kernel distance	0.709 (0.099)	0.700 (0.113)	0.703 (0.111)	0.691 (0.112)
prognostic score	0.929 (0.053)	0.878 (0.074)	0.826 (0.125)	0.845 (0.109)
	Inverse Probability Weighting: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.609 (0.135)	0.600 (0.139)	0.588 (0.161)	0.592 (0.155)
max ASMD	0.524 (0.261)	0.528 (0.259)	0.526 (0.274)	0.524 (0.270)
median ASMD	0.377 (0.252)	0.367 (0.253)	0.330 (0.268)	0.346 (0.268)
mean KS	0.479 (0.189)	0.473 (0.192)	0.437 (0.216)	0.442 (0.211)
mean t-statistic	0.609 (0.132)	0.600 (0.136)	0.590 (0.156)	0.592 (0.151)
kernel distance	0.718 (0.102)	0.714 (0.105)	0.699 (0.121)	0.698 (0.122)
prognostic score	0.952 (0.053)	0.911 (0.079)	0.886 (0.137)	0.894 (0.119)
	Subclassification: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.504 (0.197)	0.502 (0.200)	0.487 (0.208)	0.498 (0.208)
max ASMD	0.424 (0.267)	0.431 (0.265)	0.418 (0.280)	0.423 (0.278)
median ASMD	0.216 (0.249)	0.214 (0.251)	0.196 (0.247)	0.206 (0.247)
mean KS	0.120 (0.328)	0.122 (0.327)	0.092 (0.330)	0.102 (0.333)
mean t-statistic	0.537 (0.168)	0.531 (0.172)	0.529 (0.176)	0.530 (0.177)
kernel distance	0.670 (0.154)	0.671 (0.151)	0.638 (0.171)	0.648 (0.168)
prognostic score	0.794 (0.076)	0.772 (0.085)	0.765 (0.085)	0.755 (0.085)

is the average value of the five estimates. The 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 2 and Table 3, respectively.

Harder et al. [10] stated that by achieving balance, the bias in the estimated causal treatment effect due to measured covariates can be reduced. Table 2 shows that the balance metric based on kernel distance has the largest correlation with the absolute bias in the estimated casual effect, compared to the commonly used ASMD, KS statistic and t-statistic. It 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 in most cases.

To be noticed, Stuart et al. [28] proposed a new balance measure based on the prognostic score, which involves the modeling of the outcome model in the control group. In our simulation, we fit a linear parametric prognostic model with all the available covariates ($Z_1 - Z_9$) and find it shows superior performance to any method we tried. The main reason is that we fit a parametric outcome model to obtain the prognostic score, which falls into the same class of models for the true outcome according to this simulation setup. A key difference between our proposed balance metric and the prognostic score is that the former is a model-free balance metric.

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 IPW/matching/subclassification to estimate ATT and rely on a particular balance metric to select the optimal propensity score model and report ATT estimates using the chosen propensity score model. We display the bias and the standard deviation of the

Table 4: Bias and standard deviation of the estimated ATT.

	Matching: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.310 (0.787)	0.386 (1.134)	0.997 (0.604)	1.149 (0.962)
max ASMD	0.273 (0.802)	0.362 (1.161)	0.960 (0.624)	1.135 (0.975)
median ASMD	1.617 (1.822)	1.804 (2.096)	2.027 (1.577)	2.485 (2.049)
mean KS	0.686 (1.308)	0.877 (1.620)	1.279 (1.060)	1.486 (1.441)
mean t-statistic	0.302 (0.785)	0.362 (1.133)	0.982 (0.598)	1.132 (0.966)
kernel distance	0.266 (0.763)	0.341 (1.072)	0.927 (0.595)	1.075 (0.907)
prognostic score	0.096 (0.369)	0.186 (0.699)	0.834 (0.269)	0.935 (0.419)
	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	0.613 (0.598)	0.740 (0.798)	1.273 (0.512)	1.542 (0.747)
prognostic score	0.269 (0.415)	0.373 (0.640)	1.069 (0.311)	1.205 (0.450)
	Subclassification: Mean (SD)			
	Outcome A	Outcome B	Outcome C	Outcome D
mean ASMD	0.648 (0.848)	0.759 (1.146)	1.288 (0.652)	1.525 (1.022)
max ASMD	0.605 (0.830)	0.709 (1.131)	1.265 (0.638)	1.488 (1.037)
median ASMD	1.961 (1.724)	2.209 (1.975)	2.400 (1.459)	2.935 (1.958)
mean KS	1.376 (1.435)	1.589 (1.715)	1.887 (1.185)	2.265 (1.605)
mean t-statistic	0.600 (0.787)	0.705 (1.082)	1.246 (0.600)	1.466 (0.961)
kernel distance	0.643 (0.607)	0.766 (0.846)	1.270 (0.487)	1.528 (0.730)
prognostic score	0.492 (0.653)	0.575 (0.984)	1.186 (0.535)	1.413 (0.875)

estimated causal effect in Table 4, which shows the proposed balance metric can lead to the least biased and variable estimation of the causal effects for matching and inverse probability weighting, compared to other model-free balance metrics. For subclassification, kernel distance based estimator has larger bias, but its variance is the smallest, even compared to prognostic score based method.

To show the advantage of being model-free, we create another scenario where the true outcome model is as follows (the rest of the conditions are the same as before):

$$\text{Outcome Model E:} \quad Y = \beta_0 + \beta_1|Z_1| + \beta_2 Z_2 + \beta_3 Z_3 + \gamma T.$$

For the prognostic score based method, we still fit a linear parametric model with all the available covariates to obtain the prognostic score. We record the average value and the standard deviation of Pearson and Spearman's rank correlation coefficient in Table 5 based on the IPW procedure. We also report the bias and the standard deviation of the estimated ATT. As shown in the table, the prognostic score based balance metric fails completely by yielding negative correlations with the bias because it did not capture the absolute sign in the outcome model. The kernel distance shows great improvement to it, although the correlations are close to zero.

5 Data application

In this section, we are going to apply the proposed methodology to a longitudinal study: the Early Dieting in Girls Study [29]. 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 [30]. It has been shown in the literature

Table 5: Performance for Outcome Model E in Which Prognostic Score Fails.

	Pearson correlation (SD)	Spearman's correlation (SD)	Bias (SD)
mean ASMD	0.004 (0.449)	0.001 (0.435)	0.366 (0.162)
max ASMD	−0.109 (0.429)	−0.129 (0.400)	0.375 (0.181)
median ASMD	0.142 (0.295)	0.155 (0.297)	0.331 (0.145)
mean KS	0.036 (0.426)	0.027 (0.418)	0.369 (0.164)
mean t-statistic	0.002 (0.449)	−0.001 (0.436)	0.366 (0.161)
kernel distance	0.072 (0.454)	0.066 (0.440)	0.375 (0.159)
prognostic score	−0.310 (0.632)	−0.327 (0.595)	0.416 (0.106)

that mothers may influence daughters' dieting by endorsing dieting themselves and by providing information and encouragement about dieting (e. g., [31–33]). 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.

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 6, 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 remove confounding, we apply a one-to-one matching with replacement using genetic matching [34], which is a state-of-art matching procedure. In genetic matching, the generalized Mahalanobis distance (GMD) between subject *i* and subject *j* is calculated as

$$GMD(Z_i, Z_j, W) = \sqrt{(Z_i - Z_j)'(S^{-1/2})'WS^{-1/2}(Z_i - Z_j)}, \tag{9}$$

where S is sample covariance matrix of Z. W is the diagonal weight matrix where the ith diagonal 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 [5]. In Table 7, "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 6: 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.

Table 7: Causal Log Odds Ratio 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

As discussed in Diamond and Sekhon [34], 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 [35, 36]. Therefore, the covariates we aim to balance are those in Set 2 and Set 3 listed in Table 6. 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 7. 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. This is probably due to the fact that, by minimizing the kernel distance, we not only minimize the discrepancy in the finite moments of the covariates but the whole distributions so we put more constraints on the optimization process. Therefore, it leads to the most stable matching procedure. 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 [37], where mother's overall weight concern is treated as a continuous treatment variable.

To evaluate the goodness of the matching procedures, we also report the kernel distance for each matched sample. As shown in Table 7, the genetic matching procedure by minimizing the kernel distance yields the smallest kernel distance value as expected. 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 commonly used cut-off value.

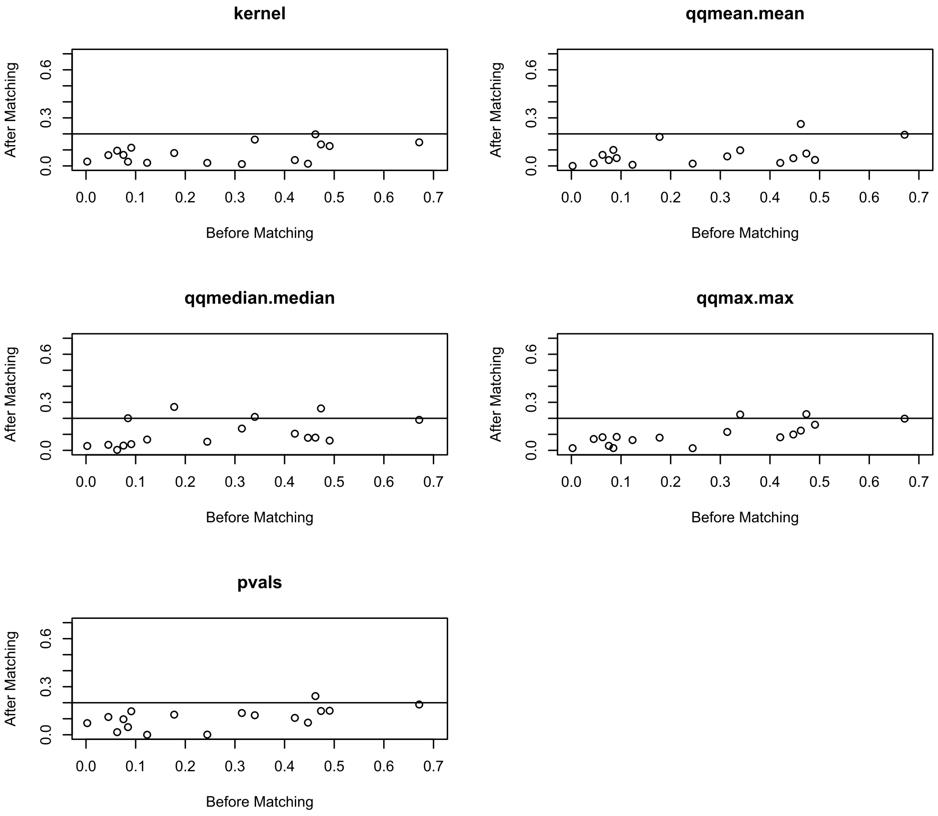

Figure 1: ASMD values before and after matching for different balance metrics in genetic matching.

Based on Gretton et al. ([38], Corollary 9), we can suggest a cut-off value for the kernel distance, which is $\sqrt{2K/m}(1+\sqrt{2\log\alpha^{-1}})$, where m 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, 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 ($\alpha = 0.05, 0.2, 0.5$), 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. It is well acknowledged that when evaluating balance, we are only concerned about whether the balance achieved in the sample is good enough to make causal inferences [4, 39]. Therefore, it is important to develop a cut-off value that does not depend on the sample size. Some future work may involve a rigorous investigation about this issue.

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.

In the data application, we employ a state-of-the-art matching approach called genetic matching to remove selection bias. We found genetic matching by minimizing the overall kernel distance leads to the most balanced matched sample and the most stable estimator. Therefore, if genetic matching is employed, we suggest using kernel distance as the stopping criterion. To be noticed, balance still needs to be checked in the individual level because it is possible some covariates will still be imbalanced after matching. In this case, the particular covariate should be included in the outcome model or another sophisticated matching procedure should be adopted.

There are several directions that are worthy of further investigation. One would be to see if one can use (6) in a manner similar to what was done in Imai and Ratkovic [13] 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.

In the simulation study, we found prognostic score based method outperforms our kernel method in a parametric simulation setting. Following the idea of prognostic score, one potential way to improve our method is to develop a kernel distance that incorporates the association between each covariate and the outcome. For example, when we use gaussian kernel to calculate kernel distance, we can modify the gaussian kernel by

$$k(x,y;\sigma^2) = \exp\{-(x-y)'W(x-y)/\sigma^2\},$$

where $W = \text{diag}(w_1,\dots,w_p)$ is a diagonal weight matrix and w_i is proportional to the association between the ith covariate and the outcome variable. A similar idea can be found in Caruana, Chevret, Resche-Rigon, and Pirracchio [40]. Future work may investigate the performance of this weighted kernel metric.

Acknowledgment: Dr. Zhu would like to thank her Master's student, Yao Li, for running some of the simulations. The content is solely the responsibility of the authors.

Funding: The project described was supported by Award Number 430-2016-00163 from the Social Sciences and Humanities Research Council and by Grant Number RGPIN-2017-04064 from the Natural Sciences and Engineering Research Council of Canada.

Appendix A. R codes

The following R codes replicate the simulation results for inverse probability weighting in Table 2, 3 and 4 based on the outcome model A in the paper.

```
library(spatstat)
library(weights)

funKS=function(w,u,z){
cdf1=ewcdf(u[z==1],w[z==1]/sum(w[z==1]))
cdf2=ewcdf(u[z==0],w[z==0]/sum(w[z==0]))
u=sort(u)
ks=max(abs(cdf1(u)-cdf2(u)))
return(ks)
}

n=1000
repe=1000
```

```
ASMDavg=rep(NA,40)
ASMDmax=rep(NA,40)
ASMDmed=rep(NA,40)
absbias=rep(NA,40)
KSavg=rep(NA,40)
TTavg=rep(NA,40)
prog=rep(NA,40)
RK=rep(NA,40)
gamma=rep(NA,40)
corr1=matrix(NA,repe,7)
corr2=matrix(NA,repe,7)
gammaopt=matrix(NA,repe,7)

for (i in 1:repe){
# Z is the treatment, Y is the outcome and W is the covariate matrix
W=matrix(NA,n,9)
W[,1]=rnorm(n)
W[,3]=rnorm(n)
W[,4]=rnorm(n)
W[,5]=rnorm(n)
W[,8]=rnorm(n)
W[,2]=rbinom(n,1,0.5)
W[,6]=rbinom(n,1,0.5)
W[,7]=rbinom(n,1,0.5)
W[,9]=rbinom(n,1,0.5)
f=log(2)*W[,1]+log(1.4)*W[,2]+log(2)*W[,4]+log(1.4)*W[,5]+log(2)*W[,7]+log(1.4)*W[,8]
+log(1.2)*W[,2]*W[,4]+log(1.4)*W[,2]*W[,7]+log(1.6)*W[,7]*W[,8]+log(1.2)*W[,4]*W[,5]
+log(1.4)*W[,1]^2+log(1.6)*W[,7]^2
ps=exp(f)/(1+exp(f))
Z=as.numeric(runif(n)<=ps)
Y=-2.4+1.68*W[,1]+1.68*W[,2]+1.68*W[,3]+3.47*W[,4]+3.47*W[,5]+3.47*W[,6]+3*Z

pshat=matrix(NA,n,40)
pshat[,1]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2+W[,7]^2,family=”binomial”)$fitted.values
pshat[,2]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2+W[,7]^2+W[,9],family=”binomial”)$fitted.values
pshat[,3]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5],family=”binomial”)$fitted.values
pshat[,4]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,1]^2+W[,7]^2,
family=”binomial”)$fitted.values
pshat[,5]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8],family=”binomial”)$fitted.values
pshat[,6]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,9],family=”binomial”)
$fitted.values
pshat[,7]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,1]^2+W[,7]^2,family=”binomial”)$fitted.values
pshat[,8]=glm(Z~W[,1]+W[,2]+W[,5]+W[,7]+W[,8]+W[,2]*W[,7]+W[,7]*W[,8]+W[,1]^2
+W[,7]^2,family=”binomial”)$fitted.values
pshat[,9]=glm(Z~W[,1]+W[,2]+W[,7]+W[,8]+W[,2]*W[,7]+W[,7]*W[,8]+W[,1]^2+W[,7]^2,
family=”binomial”)$fitted.values
pshat[,10]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5]+W[,1]^2,
family="binomial")$fitted.values

pshat[,11]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5]+W[,1]^2+W[,9],
family="binomial")$fitted.values
pshat[,12]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5],family="binomial")
$fitted.values
pshat[,13]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,1]^2,family="binomial")$fitted.values
pshat[,14]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5],family="binomial")$fitted.values
pshat[,15]=glm(Z~W[,1]+W[,4]+W[,5],family="binomial")$fitted.values
pshat[,16]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,9],family="binomial")$fitted.values
pshat[,17]=glm(Z~W[,1]+W[,9],family="binomial")$fitted.values
pshat[,18]=glm(Z~W[,2]+W[,7]+W[,8]+W[,2]*W[,7]+W[,7]*W[,8]+W[,7]^2,family="binomial")
$fitted.values
pshat[,19]=glm(Z~W[,7]+W[,8]+W[,2]*W[,7]+W[,7]*W[,8]+W[,7]^2,family="binomial")
$fitted.values
pshat[,20]=glm(Z~W[,7]+W[,8]+W[,7]^2,family="binomial")$fitted.values

pshat[,21]=glm(Z~W[,7]+W[,8],family="binomial")$fitted.values
pshat[,22]=glm(Z~W[,7]+W[,8]+W[,3]+W[,6],family="binomial")$fitted.values
pshat[,23]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2+W[,7]^2+W[,3],family="binomial")
$fitted.values
pshat[,24]=glm(Z~W[,1]+W[,2]+W[,3]+W[,4]+W[,5]+W[,6]+W[,7]+W[,8],family="binomial")
$fitted.values
pshat[,25]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5]+W[,1]^2
+W[,3]+W[,6]+W[,3]*W[,5]+W[,3]*W[,6]+W[,6]^2,family="binomial")$fitted.values
pshat[,26]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5]+W[,1]^2
+W[,3]+W[,6]+W[,3]*W[,5]+W[,3]*W[,6]+W[,6]^2+W[,9],family="binomial")$fitted.values
pshat[,27]=glm(Z~W[,1]+W[,9],family="binomial")$fitted.values
pshat[,28]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2+W[,7]^2+W[,3]+W[,6],family="binomial")$fitted.values
pshat[,29]=glm(Z~W[,4]+W[,5]+W[,7]+W[,8]+W[,7]*W[,8]+W[,3]+W[,6]+W[,3]*W[,5]
+W[,3]*W[,6]+W[,6]^2,family="binomial")$fitted.values
pshat[,30]=glm(Z~W[,1]+W[,2]+W[,5]+W[,7]+W[,8]+W[,3]+W[,6],family="binomial")
$fitted.values

pshat[,31]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,2]*W[,4]+W[,4]*W[,5]+W[,1]^2+W[,3]+W[,6],
family="binomial")$fitted.values
pshat[,32]=glm(Z~W[,3]+W[,5]+W[,6]+W[,3]*W[,5]+W[,3]*W[,6]+W[,6]^2,family="binomial")
$fitted.values
pshat[,33]=glm(Z~W[,3]+W[,5]+W[,6]+W[,3]*W[,5]+W[,3]*W[,6]+W[,6]^2+W[,9],
family="binomial")$fitted.values
pshat[,34]=glm(Z~W[,3]+W[,6]+W[,6]^2,family="binomial")$fitted.values
pshat[,35]=glm(Z~W[,3]+W[,6]+W[,6]^2+W[,9],family="binomial")$fitted.values
pshat[,36]=glm(Z~W[,3]+W[,5]+W[,6]+W[,9],family="binomial")$fitted.values
pshat[,37]=glm(Z~W[,1]+W[,2]+W[,3]+W[,5]+W[,6],family="binomial")$fitted.values
pshat[,38]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,8]+W[,2]*W[,4]+W[,2]*W[,7]
+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2+W[,7]^2+W[,3]+W[,6]+W[,3]*W[,5]+W[,3]*W[,6]
+W[,6]^2+W[,9],family="binomial")$fitted.values
pshat[,39]=glm(Z~W[,2]+W[,7]+W[,8]+W[,2]*W[,7]+W[,7]*W[,8]+W[,4]*W[,5]+W[,1]^2
+W[,7]^2+W[,3]+W[,6]+W[,3]*W[,5]+W[,3]*W[,5]+W[,3]*W[,6]+W[,6]^2+W[,9],
family="binomial")$fitted.values
pshat[,40]=glm(Z~W[,1]+W[,2]+W[,4]+W[,5]+W[,7]+W[,3]+W[,6]+W[,9],family="binomial")
$fitted.values

for (h in 1:40){
wt=1*Z+pshat[,h]/(1-pshat[,h])*(1-Z)
gamma[h]=lm(Y~Z,weights=wt)$coef[2]
absbias[h]=abs(gamma[h]-3)

ASMD=rep(NA,9)
KS=rep(NA,9)
TT=rep(NA,9)
for (j in 1:9){
ASMD[j]=abs(mean(W[Z==1,j])-sum(W[Z==0,j]*wt[Z==0])/sum(wt[Z==0]))/sd(W[Z==1,j])
KS[j]=funKS(wt,W[,j],Z)
TT[j]=wtd.t.test(W[Z==1,j],W[Z==0,j],weight=wt[Z==1],weighty=wt[Z==0],
samedata=FALSE)$coefficients[1]
}

# Calculate mean ASMD, max ASMD, median ASMD, mean KS and mean t-statistic
ASMDavg[h]=mean(ASMD)
ASMDmax[h]=max(ASMD)
ASMDmed[h]=median(ASMD)
KSavg[h]=mean(KS)
TTavg[h]=mean(TT)

# Calculate ASMD of the prognostic score
Wnew=as.matrix(cbind(1,W))
%%*%as.vector(lm(Y[Z==0]~as.matrix(W[Z==0,]))$coef)
prog[h]=abs(mean(Wnew[Z==1])-sum(Wnew[Z==0]*wt[Z==0])/sum(wt[Z==0]))/sd(Wnew[Z==1])

# Calculate kernel distance
Tstar=1/sum(Z)*Z-wt/sum(wt[Z==0])*(1-Z)
eudis=matrix(NA,n,n)
weight=matrix(NA,n,n)
for (k in 1:n){
for (m in 1:n){
eudis[k,m]=sum((W[k,]-W[m,])^2)
weight[k,m]=Tstar[k]*Tstar[m]
}
}
RK[h]=sqrt(sum(exp(-eudis/median(as.vector(eudis)))*weight))
}

# Calcuate Pearson correlation between the balance statistic and
absolute bias
corr1[i,]=apply(cbind(ASMDavg,ASMDmax,ASMDmed,KSavg,TTavg,prog,RK), absbias,
MARGIN=2,FUN=cor)
# Calcuate Spearman's correlation between the balance statistic and
absolute bias
corr2[i,]=apply(cbind(ASMDavg,ASMDmax,ASMDmed,KSavg,TTavg,prog,RK),absbias,
MARGIN=2,FUN=cor,method=”spearman”)
# Find the gamma estimate based on the optimal propensity score model
gammaopt[i,]=gamma[apply(cbind(ASMDavg,ASMDmax,ASMDmed,KSavg,TTavg,prog,RK),
MARGIN=2,FUN=which.min)]
}
```

References

1. Neyman J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Rocz Nauk Rolniczych. 1923;10:1–51.
2. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.
3. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
4. Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal. 2007;15:199–236.
5. Sekhon J. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw. 2011;42.
6. Iacus S, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assoc. 2011;106:345–61.
7. Belitser S, Martens E, Pestman W, Groenwold R, Boer A, Klungel O. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf. 2011;20:1115–29.
8. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med. 2014;33:1685–99.
9. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34:3661–79.
10. Harder V, Stuart E, Anthony J. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods. 2010;15:234–49.
11. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods. 2004;9:403.
12. Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Polit Anal. 2011. mpr025.
13. Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc, Ser B, Stat Methodol. 2014;76:243–63.
14. Hazlett C. Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. 2015. Available at SSRN 2746753.
15. Xie Y, Zhu Y, Cotton CA, Wu P. A model averaging approach for estimating propensity scores by optimizing balance. Stat Methods Med Res. 2017. https://doi.org/10.1177/0962280217715487.
16. Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. New York: Cambridge University Press; 2015.
17. Holland P. Statistics and causal inference. J Am Stat Assoc. 1986;81:945–60.
18. Zolotarev V. Probability metrics. Theory Probab Appl. 1983;28:264–87.
19. Rachev S, Klebanov L, Stoyanov S, Fabozzi F. The methods of distances in the theory of probability and statistics. Springer; 2013.
20. Wahba G. Spline models for observational data. vol. 59, SIAM; 1990.
21. Berlinet A, Thomas-Agnan C. Reproducing kernel Hilbert spaces in probability and statistics. Springer; 2011.
22. Steinwart I, Hush D, Scovel C. An explicit description of the reproducing kernel Hilbert spaces of gaussian rbf kernels. IEEE Trans Inf Theory. 2006;52:4635–43.
23. Sriperumbudur B, Fukumizu K, Gretton A, Schölkopf B, Lanckriet G, et al.. On the empirical estimation of integral probability metrics. Electron J Stat. 2012;6:1550–99.
24. Bump D. Automorphic forms and representations. Cambridge: Cambridge University Press; 1997.
25. Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GR. Hilbert space embeddings and metrics on probability measures. J Mach Learn Res. 2010;11:1517–61.
26. Adams R. Sobolev spaces. New York: Academic Press; 1975.
27. Austin P, Grootendorst P, Anderson G. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007;26:734–53.
28. Stuart E, Lee B, Leacy F. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol. 2013;66:S84–90.

29. Fisher JO, Birch L. Eating in the absence of hunger and overweight in girls from 5 to 7 y of age. Am J Clin Nutr. 2002;76:226–31.
30. Sinton M, Birch L. Weight status and psychosocial factors predict the emergence of dieting in preadolescent girls. Int J Eat Disord. 2005;38:346–54.
31. Benedikt R, Wertheim E, Love A. Eating attitudes and weight-loss attempts in female adolescents and their mothers. J Youth Adolesc. 1998;27:43–57.
32. Birch L, Fisher J. Mothers' child-feeding practices influence daughters' eating and weight. Am J Clin Nutr. 2000;71:1054–61.
33. Neumark-Sztainer D, Bauer K, Friend S, Hannan P, Story M, Berge J. Family weight talk and dieting: how much do they matter for body dissatisfaction and disordered eating behaviors in adolescent girls? J Adolesc Health. 2010;47:270–6.
34. Diamond A, Sekhon J. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat. 2013;95:932–45.
35. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25:1–21.
36. Zhu Y, Schonbach M, Coffman DL, Williams JS. Variable selection for propensity score estimation via balancing covariates. Epidemiology. 2015;26:e14–5.
37. Zhu Y, Coffman D, Ghosh D. A boosting algorithm for estimating generalized propensity scores with continuous treatments. J Causal Inference. 2015;3:25–40.
38. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. J Mach Learn Res. 2012;13:723–73.
39. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc, Ser A, Stat Soc. 2008;171:481–502.
40. Caruana E, Chevret S, Resche-Rigon M, Pirracchio R. A new weighted balance measure helped to select the variables to be included in a propensity score model. J Clin Epidemiol. 2015;68:1415–22.