Journal of Machine Learning Research 25 (2024) 1-56 Submitted 11/21; Revised 2/24; Published 5/24

A Kernel Test for Causal Association via Noise Contrastive Backdoor Adjustment

Robert Hu ROBYHU@AMAZON.CO.UK
Amazon

Dino Sejdinovic DINO.SEJDINOVIC@ADELAIDE.EDU.AU
School of Computer and Mathematical Sciences
University of Adelaide
Adelaide 5005, Australia

Robin J. Evans EVANS@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford OX1 3LB, UK

Editor: Silvia Chiappa

Abstract

Causal inference grows increasingly complex as the dimension of confounders increases. Given treatments X, outcomes Y, and measured confounders Z, we develop a non-parametric method to test the *do-null* hypothesis that, after an intervention on X, there is no marginal dependence of Y on X, against the general alternative. Building on the Hilbert-Schmidt Independence Criterion (HSIC) for marginal independence testing, we propose backdoor-HSIC (bd-HSIC), an *importance weighted* HSIC which combines *density ratio estimation* with kernel methods. Experiments on simulated data verify the correct size and that the estimator has power for both binary and continuous treatments under a large number of confounding variables. Additionally, we establish convergence properties of the estimators of covariance operators used in bd-HSIC. We investigate the advantages and disadvantages of bd-HSIC against parametric tests as well as the importance of using the do-null testing in contrast to marginal or conditional independence testing. A complete implementation can be found at `https://github.com/MrHuff/kgformula`.

Keywords: causal inference, noise contrastive estimation, kernel methods, backdoor adjustment, hsic, observational data

1. Introduction and Related Work

Modern causal inference often considers very large data sets with many observed confounding variables that may have vastly different properties. These settings are considered in a wide range of applications where randomized controlled trials are not always readily available: these include epidemiology (Rothman and Greenland, 2005), brain imaging (Castro et al., 2020), retail (Moriyama and Kuwano, 2021), and entertainment platforms (Dawen Liang and Blei, 2020). The inference setting is often complex, with high dimensional confounding variables needing to be accounted for. In such complex settings, non-parametric inference schemes that answer a simple query of causal association from observational data

are needed as an initial step before more sophisticated causal relationships can be established.

G-computation (Robins, 1986) is a classical method for estimating a causal effect from observational studies involving variables that are both mediators and confounders. Its popularity persists to this day (Daniel et al., 2013; Keil et al., 2020), because it allows one to test for a non-null causal effect using a variety of postulated models. The causal effect can also be identified for models fulfilling the so-called *backdoor criterion* (Pearl, 2009) and *ignorability assumptions* (Rosenbaum and Rubin, 1983).

In this paper, we propose a non-parametric approach to testing for the presence of a causal effect. In the Reproducing Kernel Hilbert Space (RKHS) and machine learning literature, the Hilbert-Schmidt Independence Criterion (HSIC) introduced by Gretton et al. (2005) is a widely used approach to non-parametric testing of independence. As the HSIC has good power properties and it is applicable to multivariate settings as well as to random variables taking values in generic domains, we use it as a foundation in this paper. Using g-computation principles, we introduce an extension of HSIC that can be applied to causal association testing.

Variations of HSIC have been proposed to test for conditional independence (Doran et al., 2014; Zhang et al., 2011; Honavar and Lee, 2017; Strobl et al., 2019) and while testing for conditional independence and causal association is different, some of the techniques used for conditional independence testing can be carried over to causal association. Kernel methods have been applied to tests for causal association in the binary treatment case in Muandet et al. (2021), and for average treatment effect estimation by Singh et al. (2023). In this paper, we will extend the approach of Muandet et al. to a general treatment setting. This extension is challenging as instead of modeling propensity scores, we work with general conditional densities and as such need to consider a suite of density ratio estimation techniques. We hope to bring further insights on how MMD-based approaches proposed in Singh et al. (2023) can be used for hypothesis testing.

Non-parametric tests for causal association generally come with the additional difficulty of correctly simulating the null distribution through permutations, as direct permutations often break confounding relationships as we show in Section 3.3. In this paper, we also aim to give a formal treatment on the permutation aspects of non-parametric kernel tests and expand upon ideas in Rosenbaum (1984) to consider cases beyond binary treatment.

We summarize our contributions as follows:

1. We introduce bd-HSIC, which is derived analogously to HSIC but instead uses importance weighted covariance operators, and further establish the convergence properties of the corresponding estimators in the causal setting, coupled with a novel optimization strategy to improve effective sample size.

2. We provide a novel permutation strategy for bd-HSIC that yields a permutation test with theoretically correct size for arbitrary treatment types.

3. We demonstrate that bd-HSIC has correct size and good power for different types of treatments and a large number of confounders when testing the *do-null*.

4. We analyze bd-HSIC by providing ablation studies and characterize under which circumstances it becomes invalid.

The rest of the paper is organized as follows: Section 2.2 describes the problem setting and provides a background on HSIC, Section 3 presents bd-HSIC and establishes convergence properties of the associated estimators, Section 4 presents additional details on the estimation procedure of bd-HSIC, Section 5 provides the experimental results and we conclude the paper in Section 6.

2. Background

2.1 Terminology

We start by reviewing some terminology commonly used in causal inference:

Outcome: Intuitively represents the outcome of interest possibly caused by the treatment. It is denoted by Y and could be continuous such as blood pressure, or binary if it represented as recovery or not from disease.

Treatment: Treatments intuitively refer to the variables whose effect on the outcome Y we are trying to infer. We denote *treatments* by X and in the case where treatment is placebo-controlled we can consider $X \in \{0, 1\}$ representing control and active treatment respectively. In general however, treatment could be a random variable taking values in continuous or multivariate domains.

Observed confounding variables: Intuitively represents patient characteristics that may affect both treatment and outcome. We denote these variables with Z and these could be continuous variables such as height and weight, or binary variables such as sex.

Assignment: In many confounded settings, certain patient characteristics (Z) affect *treatment assignment*. As an example, one can imagine socioeconomic factors having an effect on the availability of medical treatment. This creates a bias in treatment assignment which needs to be adjusted for.

2.2 Setup

Consider a situation where we observe treatments X, outcomes Y and potential confounders Z defined on measurable spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively. We assume these quantities are observed as $\{(z_i, x_i, y_i)\}_{i=1}^n \sim p$, where p is some joint probability density on the product space $\mathcal{Z} \times \mathcal{X} \times \mathcal{Y}$. We are interested in establishing a causal relationship between treatments X and outcomes Y, meaning that we manage to isolate whether there precisely exists a dependency between X and Y whilst holding all pre-treatment variables constant. In ideal circumstances, we would not have any confounders and such a relationship can be established straightforwardly, using e.g. regression methods. However, such circumstances are unusual outside randomized trials and would require

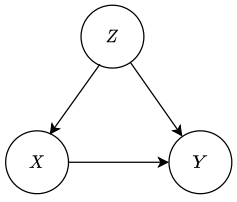


FIGURE 1. Graph showing the relationship between treatments X, outcomes Y, and confounders Z

that the treatment is assigned to units by some exogenous process. In the more common case of observational studies, dependencies among X, Y, Z are often depicted as in Figure 1. The existence of confounders Z complicates establishing the causal relationship between X and Y, as they introduce dependence between X and Y which is difficult to disentangle from the postulated causal effect.

A motivating real-world problem In development economics, it is of great importance to establish causes of infant mortality (Ensor et al., 2010). The causes may often have a non-linear association with the mortality while being confounded by circumstantial factors such as the socio-economical background of the parents and the medical history of the mother.

In this setting, it is also important to test for *distributional differences* different policies might induce, as tests based on average treatment effect may fail to capture such differences as discussed in (Muandet et al., 2021; Fawkes et al., 2022). As the mean embedding framework offers a natural way to non-parametrically quantify distributional differences, kernel-based test statistics will be our preferred strategy. We will illustrate throughout the paper the importance of a non-parametric test that is able to capture non-linear dependencies for different treatments under a large number of confounders.

Definition 1. *(do-null hypothesis) Let*

$$p(y \mid do(x)) := \int p(y \mid x, z)\, p(z)\, dz$$

where in general $p(y|x) \neq p(y \mid do(x))$ since X and Z may be dependent. We are interested in testing if the interventional distribution $Y \mid do(X = x)$ does not depend on the value of the treatment variable X. We refer to this hypothesis as a do-null hypothesis*, which can be stated as $Y \perp X \mid do(X = x)$. In terms of distributions, we can consider*

$$H_0 : p(y \mid do(x)) = p^*(y) \tag{1}$$

as our null hypothesis versus the general alternative, where p^* is an arbitrary distribution that does not depend upon the value of X. We consider H_0 for all values of $do(x)$ and the observational regime. Note that our null hypothesis does not *imply that $p(y \mid do(x)) = p(y)$, which is why we introduce p^*.*

For a remark on why we need p^*, see Appendix A.

In this paper, we will introduce a test statistic for the hypothesis (1), but first, we review HSIC (Gretton et al., 2005), which serves as the foundation for our contribution.

2.3 Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion (HSIC) is a powerful non-parametric test of independence for high-dimensional data. It considers the problem of empirically establishing whether there is any form of departure from independence between two random variables taking values on generic domains.

Definition 2 (Marginal Independence Testing). *Let P_{xy} be a Borel probability measure defined on a domain $\mathcal{X} \times \mathcal{Y}$ and let P_x and P_y be the respective marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$. Given an i.i.d. sample $(X, Y) = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of size m drawn according to P_{xy}, does P_{xy} factorize as $P_x P_y$? We usually consider the null hypothesis to be*

$$H_0 : P_{xy} = P_x P_y$$

against the general alternative

$$H_1 : P_{xy} \neq P_x P_y.$$

Since we do not have access to P_{xy}, P_x, or P_y, we need to estimate or represent these distributions through either parametric or non-parametric means. A convenient way for representing distributions is to use the RKHS formalism (Scholköpf and Smola, 2001).

HSIC can intuitively be understood as a covariance between RKHS representations of random variables.

Definition 3 (Reproducing Kernel Hilbert Spaces). *Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is called a* reproducing kernel Hilbert space *endowed with dot product $\langle \cdot, \cdot \rangle$ if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:*

1. *k has the reproducing property*

$$\langle f, k(x, \cdot) \rangle = f(x), \quad \forall f \in \mathcal{H}, x \in \mathcal{X};$$

2. *k spans $\mathcal{H}$, that is, $\mathcal{H} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$ where the bar denotes the completion of the space.*

We generally refer to the function k as a *kernel*. For certain choices of k, the corresponding RKHS is *characteristic*, i.e. the mapping of a probability measure μ, $\mu \mapsto \int_X k(x, \cdot) d\mu(x)$ is injective, i.e. μ is mapped to a unique element in the RKHS. We refer to Sriperumbudur et al. (2011) for more details. This property is very practical, as it allows us to embed probability distributions into the RKHS and calculate their expectations based on observations. These embeddings are called kernel mean embeddings, see Muandet et al. (2017) for a thorough exposition.

Definition 4. *Let $\mathcal{X}$ be a measurable space and let $\mathcal{H}_\mathcal{X}$ be an RKHS on $\mathcal{X}$ with kernel k. Let P be a Borel probability measure on $\mathcal{X}$. An element $\mu_x \in \mathcal{H}_\mathcal{X}$ such that $\mathbb{E}_{x \sim P}[f(x)] = \langle f, \mu_x \rangle$, $\forall f \in \mathcal{H}_\mathcal{X}$ is called the* ***kernel mean embedding*** *of P in $\mathcal{H}_\mathcal{X}$, where $\mu_x : P \mapsto \int_\mathcal{X} k(x, \cdot) dP(x)$.*

A sufficient condition for the existence of a kernel mean embedding is that $\mathbb{E}_{x \sim P}[\sqrt{k(x, x)}] < \infty$, which is satisfied for, e.g. bounded kernel functions.
Given some observations $\{x_i\}_{i=1}^n \sim P$, the empirical mean embedding of P is estimated as:

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot).$$

Kernel mean embeddings intuitively allow us to estimate expectations under each of p_{xy}, p_x, and p_y. In order to test for marginal dependence, we consider a test statistic based on $\mathrm{Cov}[f(X), g(Y)]$, where f, g are arbitrary continuous functions evaluating random variables X, Y. Instead of picking individual functions f,g, we consider the representation of their covariance using the RKHS.

Definition 5. *Let (X, Y) be a pair of random variables defined on $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ be RKHSs on $\mathcal{X}$ and $\mathcal{Y}$, respectively. An operator $C_{X,Y} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ such that*

$$\langle g, C_{X,Y} f \rangle = \mathrm{Cov}[f(X), g(Y)], \ \forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}$$

*is called a **cross-covariance operator** of X and Y. We note that $C_{X,Y}$ is the property of the joint distribution of pair (X, Y). We denote it simply by C when there is no ambiguity. The Hilbert-Schmidt Independence Criterion (HSIC) is then defined as the squared Hilbert-Schmidt (HS) norm of C, i.e.*

$$\mathit{HSIC}(X, Y) := \|C\|^2_{\mathrm{HS}} = \sum_i \sum_j \langle C u_i, v_j \rangle^2$$

with u_i and v_j being orthogonal basis of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ respectively.

It is readily shown via reproducing property that C can be written as

$$C := \mathbb{E}\left[(k(X, \cdot)) \otimes (l(Y, \cdot))\right] - \mathbb{E}\left[k(X, \cdot)\right] \otimes \mathbb{E}\left[l(Y, \cdot)\right],$$

where k, l are kernels of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ respectively, and $\otimes$ denotes the outer product. An alternative view of HSIC is that it measures the squared RKHS distance between the kernel mean embedding of P_{xy} and $P_x P_y$. For sufficiently expressive kernels (Sriperumbudur et al., 2011), this distance is zero if and only if X and Y are independent. To see that C indeed is a Hilbert-Schmidt operator, we refer to Muandet et al. (2017).
Given a sample $\{(x_i, y_i)\}_{i=1}^n$ from the joint distribution P_{xy}, an estimator[1] of HSIC is given by:

$$\begin{aligned}\widehat{\mathrm{HSIC}}(X, Y) = &\frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) l(y_i, y_j) + \frac{1}{n^2} \sum_{i,j=1} k(x_i, x_j) \frac{1}{n^2} \sum_{k,l=1}^n l(y_k, y_l) \\ &- \frac{2}{n^3} \sum_{i,j,k=1}^n k(x_i, x_j) l(y_i, y_l).\end{aligned}$$

The estimator above serves as the test statistic, for complete derivation, we refer to Gretton et al. (2005). To estimate its distribution under the null hypothesis, we resort to repeatedly permuting y_i's to obtain $\left\{\left(x_i, y_{\pi(i)}\right)\right\}_{i=1}^n$ where π is a random permutation, and recomputing HSIC on this permuted data set. We note that the asymptotic null distribution of HSIC has a complicated form (Zhang et al., 2017), and it is hence standard practice to use a permutation approach to approximate it. For more details on HSIC, we refer to (Gretton et al., 2005).

1. This is the most commonly used, biased estimator of HSIC. An unbiased estimator also exists, cf. Song et al. (2007)

2.4 Difference between the do-null, marginal independence and conditional independence

While the do-null intuitively bears many similarities to marginal independence and conditional independence, we briefly illustrate the difference between these independencies and provide an experimental demonstration that conditional independence tests (RCIT) and marginal independence tests (HSIC) cannot be used to test for the do-null.

2.4.1 The do-null is not marginal independence $X \perp Y$.

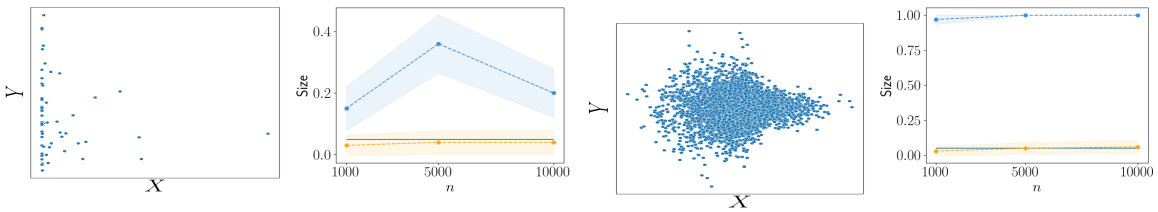
We contrast the do-null with marginal independence $X \perp Y$, by constructing a data set where the do-null is true, but there is marginal dependence between X and Y. We illustrate the dependency in Figure 2a.

We further plot size ($\alpha = 0.05$) against sample size when applying HSIC and bd-HSIC (true weights) in Figure 2b. HSIC is not calibrated under the do-null.

This can further be interpreted as equality in distribution of Y for all values $do(x)$ excluding the observational regime, meaning our hypothesis does not test for marginal independence between X and Y.

2.4.2 The do-null is not conditional independence $X \perp Y \mid Z$.

We similarly contrast the do-null with conditional independence $X \perp Y \mid Z$. We simulate a data set such that there is a conditional dependence $X \not\perp Y \mid Z$ while the do-null is true. See Figure 2c for an illustration of the dependency between X and Y. We apply the "RCIT" method, a kernel-based conditional independence test proposed in Strobl et al. (2019) and demonstrate in Figure 2d that it rejects almost all the time when applied to the data under the do-null.



(A) Marginal dependency but true do-null (B) bd-HSIC(true weights): vs HSIC: (C) Conditional dependency but true do-null (D) bd-HSIC(true weights): vs RCIT:

Figure 2. Difference between the do-null, marginal independence and conditional independence. See Appendix F.2 on how the data in Figure 2a and Figure 2c is generated.

3. Backdoor-HSIC

Our proposed method has two parts, a weighted HSIC test statistic and a density ratio estimation procedure. In this section, we introduce the test statistic, which we term *backdoor-HSIC* (bd-HSIC).

3.1 Overview

3.1.1 The do-operator and identifiability

We start with reviewing the meaning of the do-operator and how it lays the foundation for bd-HSIC. In the remainder of the paper, we will assume that all relevant probability distributions admit densities.

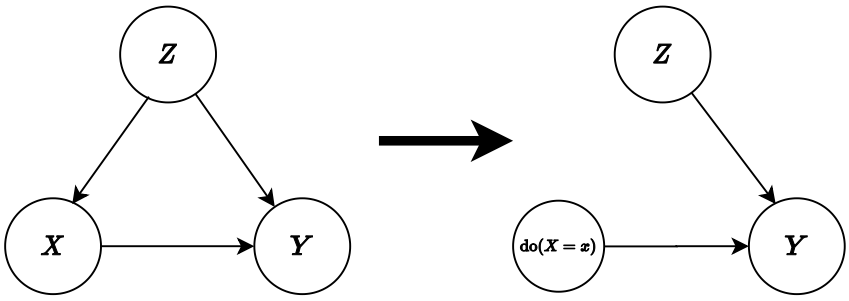


Figure 3. Illustration of the interventional distribution $p(y \mid do(x))$

In Figure 3, we first consider the graphical representation of the problem of establishing the relationship between X and Y under observed confounders Z. We apply the do-operation (Pearl, 2009) on our treatment X, in order to remove any dependency between X and Z. This adjusts for the effect of confounding on our treatment. The resulting distribution for the above graph, which we denote $p(z, y \mid do(x))$, can be seen as the conditional distribution of Y, Z given X where we have made Z and X independent, but not affected the conditional distribution of Y given Z, X. This distribution can then be used to understand the *causal* relationship between X and Y.

We want to ensure the causal effect adjusted for confounders Z is identifiable for the graph in Figure 3. We revisit the *backdoor criterion* presented by Pearl (2009).

Definition 6. *A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (A, B) in a DAG G if:*

- *no node in Z is a descendant of A; and*
- *Z blocks every path between A and B that contains an arrow into A.*

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (A, B) such that $A \in X$ and $B \in Y$.

We see that our confounders Z satisfy the backdoor criterion. By Theorem 3.3.2 in Pearl (2009) the causal effect is identifiable as

$$\begin{aligned} p^*(y|x) = p(y|do(x)) &= \int p(y|do(x), z)p(z \mid do(x))dz \\ &= \int p(y|\underbrace{x, z}_{\text{Only causal association between } x \text{ and } y})p(z \mid do(x))dz \\ &= \int p(y|x, z)p(\underbrace{z}_{\text{Dependency between } x \text{ and } z \text{ gone}})dz. \end{aligned}$$

Given some arbitrary density $p^*(x)$, this defines a joint density $p^*(z, x, y) = p(z)p^*(x)p(y|x, z)$.

The core idea is to calculate the HSIC between X and Y under the interventional regime, a distribution we denote by p^*. and conduct a permutation test to establish whether $p^*(y) = p^*(y \mid x)$. Since we consider the interventional distribution $p^*(y \mid x)$, care must be taken to obtain the correct mean embedding. It turns out that one can use importance weights to express the mean embedding of $p^*(y \mid x)$, where the importance weights are defined as

$$w(x, z) = \frac{p^*(x)}{p(x \mid z)}, \tag{2}$$

which defines a *density ratio*. Note that the dependence upon z complicates the necessary permutation procedure. Our method then comprises of two steps:

1. Estimate the importance weights $w(x, z)$.

2. Run a permutation test and calculate a p-value to establish whether $p^*(y) = p^*(y \mid x)$.

We present our proposed method in detail for step 2 in the remainder of this section, and step 1 together with the overall procedure in Section 4.

3.2 Estimation

We are now ready to present the first link between HSIC and $p(y \mid do(x))$, as we are interested in expectations under p^* using samples from p. We define the expectation operator $\mathbb{E}_p$ in the usual way:

$$\mathbb{E}_p\left[f(X_1, \ldots, X_k)\right] = \int \cdots \int f(x_1, \ldots, x_k)p(x_1, \ldots, x_k)\, dx_1 \ldots dx_k.$$

We would also like to calculate the expectation under the interventional distribution p^*, using samples from p. This can be done by using a weight function $w(z, x) = p^*(x)/p(x \mid z)$ so that

$$\mathbb{E}_{p^*}[f(Z, X, Y)] = \mathbb{E}_p[w(X, Z) \cdot f(Z, X, Y)].$$

Proposition 1. *Consider continuous and bounded real-valued functions f, g. The covariance between $f(X)$ and $g(Y)$ under p^* can be calculated as*

$$\begin{aligned}\mathrm{Cov}_{p^*}[f(X), g(Y)] &= \mathbb{E}_{p^*}[f(X)g(Y)] - \mathbb{E}_{p^*}[f(X)]\,\mathbb{E}_{p^*}[g(Y)] \\ &= \mathbb{E}_p[Wf(X)g(Y)] - \mathbb{E}_{p^*}[f(X)]\,\mathbb{E}_p[Wg(Y)],\end{aligned}$$

where $W = w(X, Z)$, provided that $p(X|Z) > 0, \forall X, Z$ s.t. $p^*(X) > 0$ and the integrals exist. Using these weights we can now calculate any expectation term under p^* in the covariance estimator.

Proof We show that $\mathbb{E}_{p^*}[g(Y)]$ and $\mathbb{E}_{p^*}[f(X)g(Y)]$ indeed can be calculated as $\mathbb{E}_p[Wg(Y)]$ and $\mathbb{E}_p[Wf(X)g(Y)]$ respectively. To see this, we have that

$$\begin{aligned}\mathbb{E}_p[Wf(X)g(Y)] &= \iiint f(x)g(y)\frac{p^*(x)}{p(x|z)}p(x, y, z)\,dx\,dy\,dz \\ &= \iiint f(x)g(y)p(z)p^*(x)p(y \mid z, x)\,dz\,dx\,dy \\ &= \iiint f(x)g(y)p(z)p^*(z, x, y)\,dz\,dx\,dy \\ &= \mathbb{E}_{p^*}[f(X)g(Y)].\end{aligned}$$

The case for $f(x) = 1$ then also follows. ■

In the above example, we will need to estimate importance weights w_i from observations $x_i \in \mathcal{X}$ and $z_i \in \mathcal{Z}$ such that $w_i = p^*(x_i)/p(x_i|z_i)$. Under the do-null, we have that $\mathbb{E}_p[Wf(X)g(Y)] = \mathbb{E}_{p^*}[f(X)]\mathbb{E}_p[Wg(Y)]$.

3.2.1 Choosing the marginal distribution of X

Calculating the covariance between two arbitrary functions can generally be challenging and may require parametric assumptions. This is obviously undesirable, so similarly to HSIC, we will consider cross-covariance operators, which represent the covariances between two functions of the variables. The key object of interest is the cross-covariance operator of treatments X and outcomes Y under p^* in the interventional regime, which we denote by C_{p^*}. This operator plays an analogous role to that of the cross-covariance under an observational distribution p in standard independence testing. In particular, the squared HS norm of C_{p^*} is the population HSIC under p^* and hence the size of this operator measures departure from the do-null hypothesis. Similarly to HSIC, whenever we use characteristic kernels, we have that $\|C_{p^*}\|^2 = 0 \iff p^*(y) = p(y \mid do(x))$. Of course, we are unable to estimate this quantity directly since we do not have access to samples from $p^*(y|x)$. The following immediate corollary to Proposition 1 relates C_{p^*} to expectations under p.

Corollary 1. *The cross-covariance operator C_{p^*} of X and Y under $p^*(y|x)$ satisfies*

$$\langle f, C_{p^*} g\rangle = \mathbb{E}_p[W_{p^*} f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}_p[W_{p^*} g(Y)],\ \forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}.$$

and C_{p^*} can be expressed as

$$C_{p^*} := \mathbb{E}_p[W_{p^*} k(X, \cdot) \otimes l(Y, \cdot)] - \mathbb{E}[k(X_{p^*}, \cdot)] \otimes \mathbb{E}_p[W_{p^*} l(Y, \cdot)]$$

with $W_{p^*} = \frac{p^*(X)}{p(X \mid Z)}$.

Following this corollary, we can empirically estimate C_{p^*} using the following expression:

$$\widehat{C}_{p^*} = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i k(\cdot, x_i) \otimes l(\cdot, y_i) - \left(\frac{1}{n}\sum_{j=1}^{n} k\left(\cdot, x_j^{p^*}\right)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i l\left(\cdot, y_i\right)\right), \tag{3}$$

where $\{(z_i, x_i, y_i)\}_{i=1}^n \sim p(z, x, y)$ and $\left\{x_j^{p^*}\right\} \sim p^*$. In the case where both $p^*(x_i)$ and $p(x_i|z_i)$ are known, one would simply use the "true weights" $w_i = \frac{p^*(x_i)}{p(x_i|z_i)}$. We can show that the resulting estimator (3) is consistent.

Theorem 1. *Assuming* $Var\left(\frac{p^*(X)}{p(X|Z)}\right) < \infty$, *and that* $\mathbb{E}_{X,X'}[k(X, X')] < \infty$, $\mathbb{E}_{Y,Y'}[l(Y, Y')] < \infty$, *then* $\widehat{C}_{p^*}$ *using true weights is a consistent estimator of* C_{p^*} *and satisfies*

$$\mathbb{E}\left[\|C_{p^*} - \widehat{C}_{p^*}\|_{\mathrm{HS}}^2\right] = \mathcal{O}\left(\frac{1}{n}\right).$$

Proof *See Appendix B.1.* ■

In practice, however, the true weights are rarely available and they would need to be estimated using density ratio estimation techniques, which we shall discuss in detail in Section 4. The weights estimation corresponds to estimating a function h, s.t. $\tilde{w}_i = h(x_i, z_i)$. We note that estimating h will need to be performed on *a different data set* than the one used to estimate the covariance in (3), to ensure independence between h and the samples used for testing independence. We now give a result regarding the convergence rate when density ratios are estimated, which shows how the convergence rate of the density ratio estimator affects that of the corresponding estimator of the cross-covariance operator.

Theorem 2. *Take the conditions of Theorem 1, and assume also that* $\hat{h}_n(x, z)$ *is a consistent estimator of the density ratio* $\frac{p^*(x)}{p(x \mid z)}$ *with uniform convergence rate* $\mathcal{O}(\frac{1}{n^\alpha})$ *for* $\alpha > 0$, *i.e.*

$$\lim_{n\to\infty} \sup_{x,z} \left|\hat{h}_n(x, z) - \frac{p^*(x)}{p(x \mid z)}\right| \propto \mathcal{O}\left(\frac{1}{n^\alpha}\right).$$

Then

$$\mathbb{E}\left[\left\|C_{p^*} - \widehat{C}_{p^*}\right\|_{\mathrm{HS}}^2\right] = \mathcal{O}\left(\frac{1}{n^{\min(1,\alpha)}}\right).$$

Proof *See Appendix B.2.* ■

In summary, the convergence rate for the bd-HSIC estimator using estimated weights is at worst the slower rate between the estimator and the uniform convergence rate of the weight estimates. Now that we have established how to estimate the cross-covariance operator C_{p^*}, analogously to HSIC, we will use the squared HS norm of $\hat{C}_{p^*}$ as our test statistic.

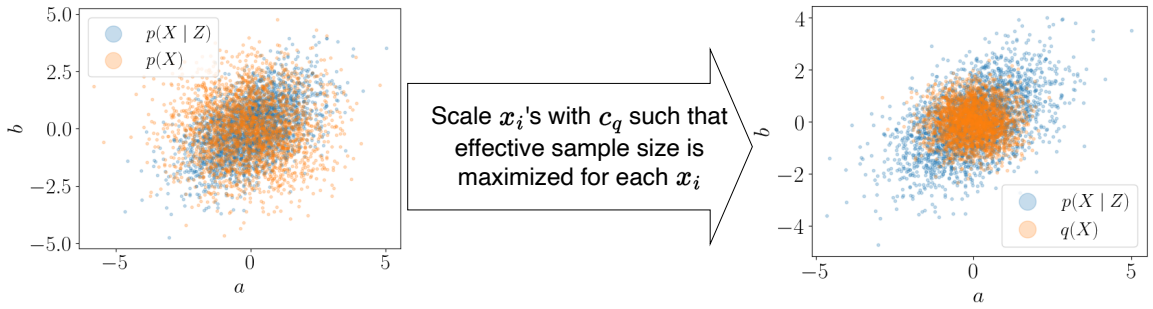


FIGURE 4. In the left example plot, $p(X)$ has samples where $p(X \mid Z)$ has little support. By constructing $p^*(x)$, samples from $p^*(x)$ have support where $p(X \mid Z)$ has support, which yields a higher effective sample size. In this example, we consider $p(X)$ and $p(X \mid Z)$ a 2d distribution with observations on the form $x = (a, b)$. The axes represent the values for a and b.

Remark 1. *To the best of our knowledge, no density ratio estimator we consider in this work provides any uniform convergence guarantees. Theorem 2 only gives an indicative convergence rate if such density ratio estimation guarantees could be established. In the absence of such bounds, consistency will be hard to guarantee without other strong assumptions. The result can be further refined to settings where uniform convergence is not assumed for the density ratio estimator, however, we leave this to future work.*

Proposition 2. *Let $\circ$ denote the element-wise matrix product and $\mathbf{K}_{++}$ denote the sum of all elements in the matrix $\mathbf{K}$. The squared HS norm of the estimator in* (3) *is given by*

$$\|\widehat{C}_{p^*}\|^2_{HS} = \frac{1}{n^2}\tilde{\mathbf{w}}^\top (\mathbf{K} \circ \mathbf{L})\, \tilde{\mathbf{w}} + \frac{1}{n^4}(\mathbf{K}^{X_{p^*},X_{p^*}})_{++}(\mathbf{L} \circ \tilde{\mathbf{W}})_{++} - \frac{2}{n^3} \cdot \tilde{\mathbf{w}}^\top \left(\mathbf{K}^{X,X_{p^*}} \mathbf{1}_n \circ \mathbf{L}\tilde{\mathbf{w}}\right) \quad (4)$$

where $\tilde{\mathbf{w}} = [h(x_1, z_1), \ldots, h(x_n, z_n)]$, $\tilde{\mathbf{W}} = \tilde{\mathbf{w}}^\top \tilde{\mathbf{w}}$, $\mathbf{K} = [k(x_i, x_j)]_{i,j=1}^n$, $\mathbf{L} = [l(y_i, y_j)]_{i,j=1}^n$, $\mathbf{K}^{X_{p^*},X_{p^*}} = [k(x_i^{p^*}, x_j^{p^*})]_{i,j=1}^n$, $\mathbf{K}^{X,X_{p^*}} = [k(x_i, x_j^{p^*})]_{i,j=1}^n$ and $\mathbf{1}_n$ is a vector of ones with length n. **Proof** *See Appendix C.* ■

Equation 4 can be viewed as a weighted version of HSIC. We henceforth will refer to this as *bd-HSIC* and use it as our test statistic. We note that a weighted form of HSIC has previously been considered in a different context – when testing for independence on right-censored data (Rindt et al., 2020).

3.2.2 The choice of the p^*-marginal

Using $p^*(x)$ vs $p(x)$ In general, we would expect $p^*(x)$ to provide a higher effective sample size of w's if chosen appropriately. To maximize effective sample size we can choose $p^*(x)$ such that samples $x_i^{p^*} \sim p^*(x)$ are given by $x_i^{p^*} = c_{p^*} \cdot x_i$. For continuous x_i with mean 0, this can be viewed as scaling the variance of samples x_i. We illustrate this in Figure 4. We describe how to choose an optimal c_{p^*} for continuous densities in the paragraphs below.
Choice of c_{p^*} To find an optimal c_{p^*} for univariate and multivariate X, Z, we optimize the

effective sample size of $w_i = \frac{p^*(x_i)}{p(x_i \mid z_i)}$. We choose the effective sample size to be as large as possible to try to maximize the power of the test. Since the marginals of X, Z are generally unknown, we derive a heuristic based on Gaussian distributions.

Proposition 3. *Assuming standard normal marginals for univariate X, Z with correlation ρ. Then c_{p^*} maximizes effective sample size (ESS)*

$$ESS := \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$$

where w_i is the weight used to re-sample observations. The optimal c_{p^*} is then given by $c_{p^*} = \sqrt{1 - 2\rho^2}$, where ρ is the correlation between X and Z.

Proof See Appendix D.1. ■

Here, under the assumption of joint normality, the optimal c_{p^*} is found analytically.

Proposition 4. *Assuming $(X, Z) \sim \mathcal{N}_{m+n}(0, \Sigma)$, where we take $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}$ and assume that $\Sigma_{xx} = I_m$ and $\Sigma_{zz} = I_n$, the optimal new covariance matrix is found by maximizing the quantity*

$$\det(D) \det\left(2T^{-1} - (I_m - \Sigma_{xz}\Sigma_{zx})^{-1} - BD^{-1}B^\top\right)$$

with respect to the positive definite matrix T, where $B := (I_m - \Sigma_{xz}\Sigma_{zx})^{-1} \Sigma_{xz}$ and $D := I_n - \Sigma_{zx} (I_m - \Sigma_{xz}\Sigma_{zx})^{-1} \Sigma_{xz}$.

Proof See Appendix D.2. ■

In practice in the multivariate case, we choose $T = I_m \cdot c_{p^*}$, and perform the optimization using gradient descent with $c_{p^*} \in \mathbb{R}_{>0}$.

3.3 The importance of using the correct permutation

For HSIC, Rindt et al. (2021) shows that permutation tests indeed provide the correct size under the null. However, when permuting under confounders their Theorem 2 does not hold anymore due to the extra dependency on Z between X and Y.

To see this, consider i.i.d. samples $(z_i, x_i, y_i) \sim p$ and that H_0 is true, i.e. $p(y \mid do(x)) = p^*(y)$. If you would permute on x_i's or y_i's directly it would violate the exchangeability and hence the guarantee that the test will have the correct size α in expectation, due to the additional dependencies with z_i's. Permuting y_i's directly breaks the required dependency with z_i. Further permuting x_i's directly breaks the interventional distribution $do(X = x)$ as it is obtained by re-weighting x_i's according to $\frac{p^*(x_i)}{p(x_i|z_i)}$. Thus we have that, for a random permutation D,

$$(\underbrace{D\mathbf{x}}_{\text{permuting } x_i\text{'s}}, \mathbf{y}) \overset{d}{\neq} (\mathbf{x}, \mathbf{y})$$

	$z=0$	$z=1$
$x=0$	p	$1-p$
$x=1$	$1-p$	p

Table 1. Probabilities for $p(X = x \mid Z = z)$

and

$$(\mathbf{x}, \underbrace{D\mathbf{y}}_{\text{permuting } y_i\text{'s}}) \overset{d}{\neq} (\mathbf{x}, \mathbf{y}).$$

Hence the size of the test is not guaranteed to be at most α.

To circumvent this, we consider a similar approach to Doran et al. (2014) who permute only where the covariates Z are similar. As bd-HSIC uses the density ratio as an importance weight, we show that permuting Y, Z against X where $p(x_i \,|\, z_i)$ is similar yields an exchangeable sample.

Theorem 3. *(Correct type 1 error rate for bd-HSIC using finite permutations) Assume we have i.i.d. samples $(z_i, x_i, y_i) \sim p$ and that H_0 is true, i.e. $p(y \,|\, do(x)) = p^*(y)$. If a permutation test with finite samples is applied to bd-HSIC to test for causal association at level α, this test will reject with probability at most α if only y_i's with the same conditional density $p(x_i|z_i)$ are being permuted.*

Proof Let D' be a permutation such that only y_i's with the same conditional density $p(\cdot|z_i)$ are being permuted. As the samples are permuted within $p(\cdot|z_i)$, the following holds

$$(\mathbf{x}, \mathbf{y}) \overset{d}{=} (\mathbf{x}, \underbrace{D\mathbf{y}}_{\text{permuting } y_i\text{'s}})$$

for any permutation D. The computed test statistics (unpermuted and permuted data) are now exchangeable under the null, and thus the rankings will be uniformly distributed. ■

It might seem counter-intuitive that the weights w_i from the density ratio estimation cannot be used for the permutation test. We provide a corollary below why using such w_i's would fail.

Corollary 2. *Assuming a non-constant true density ratio weights, permuting "within" the density ratio weights w_i yields incorrect permutations.*

Proof We provide an example that breaks exchangeability when permuting within density ratio weights. Consider binary Z and X with $p = 0.5$ marginally, with Y being a clone of Z. Let $p(X = x \mid Z = z)$ be defined as in Table 1 with $p \neq 0.5$. As Y is a clone of Z, $p(X = x \mid Y = y)$ has the same probabilities as above. Under these assumptions, it is easy to see that (x, y) samples $(0, 0)$ and $(1, 1)$ (as well as $(1, 0)$ and $(0, 1)$) will have to permute since $p(X = 0 \mid Y = 0) = p(X = 1 \mid Y = 1) = p$. If one considers a situation in which the same number of pairs ($0, 0$ and $1, 1$) and ($1, 0$ and $0, 1$) appear in a sample, the

resulting conditional distribution under permutation will be $p(Y = y \mid X = x) = 0.5$ for all $x, y \in \{0, 1\}$. Since we took $p \neq 0.5$, the permuted samples are no longer exchangeable with the original sample, even though the do-null clearly holds. ■

3.3.1 MMD clustering

In practice, we are unlikely to get a group of samples with exactly the same density ratio. To approximately permute y_i's within $p(x_i \mid z_i)$, we propose a maximum mean discrepancy (MMD, Gretton et al., 2005) based k-means clustering method (MacQueen, 1967), which uses non-parametric kernel conditional density estimation. The procedure can be summarized in Algorithm 1.

Algorithm 1: Clustering using Conditional Mean Embedding

Input: Training data $\{x_i\}_{i=1}^{n_{\text{tr}}}$, labels $\{z_i\}_{i=1}^{n_{\text{tr}}}$, test data $\{z_j\}_{j=1}^{n_{\text{test}}}$, regularization parameter λ, number of clusters k

Output: Cluster assignments c_j for test data points

Training Phase:

Estimate the conditional mean embedding: $\hat{\mu}_{X|Z=z}(\cdot) = \sum_{i=1}^{n_{\text{tr}}} w_i(z) l(x_i, \cdot)$

Compute $\mathbf{w}(z) = (\mathbf{L}_Z + \lambda I)^{-1}\mathbf{l}_z$, where $\mathbf{l}_z = (l(z_1, z), \ldots, l(z_{n_{\text{tr}}}, z))^\top$.

Run k-means to get cluster centers C_k, by optimizing the MMD metric:
$\sum_i \|\hat{\mu}_{X|Z=z_i}(\cdot) - \hat{\mu}_{X|Z=C_k}(\cdot)\|^2 = \sum_{i=1}^{n_{\text{tr}}} (\mathbf{w}(z_i) - \mathbf{w}(C_k))^\top \mathbf{L}_X (\mathbf{w}(z_i) - \mathbf{w}(C_k))$.

Testing Phase: for $j = 1$ *to* n_{test} **do**

  $c_j = \min_k (\mathbf{w}(z_j) - \mathbf{w}(C_k))^\top \mathbf{L}_X (\mathbf{w}(z_j) - \mathbf{w}(C_k))$.

end

Return: Cluster assignments c_j for test data points

One permutes y_j's within assigned clusters to get a D' permutation. To select the optimal number of clusters, we maximize the silhouette score (Rousseeuw, 1987) over a fixed number of clusters.

Remark 2. *We note that MMD-clustering procedure uses Conditional Mean Embedding (CME) estimators (since the true $p(x \mid z)$ is unknown). Li et al. (2023) show these estimators to be consistent with rate $\mathcal{O}(\frac{\log n}{n})$, under realistic smoothness assumptions. It is an interesting avenue for further research to investigate if the clustering obtained using estimated CMEs can be related to the clustering obtained using the true CMEs in the large sample limit.*

3.4 Consistency of bd-HSIC

So far we have provided results on correct type 1 errors for permutation tests. To prove the asymptotic consistency of bd-HSIC, it suffices to show that Lemma 1 in Rindt et al. (2021) holds for bd-HSIC, assuming correctly estimated weights and access to $p(x_i|z_i)$. Consistency then follows from Theorem 3 in Rindt et al. (2021).

Proposition 5. *Let ψ be a random permutation of y_i's such that they only are permuted within groups that have the same $p(x_i \mid z_i)$. We write $\|\widehat{C}_{p^*}\|^2_{HS}(\psi)$ to denote the HS-norm of $\widehat{C}_{p^*}$, under the permutation ψ to all y_i. Then*

$$\|\widehat{C}_{p^*}\|^2_{HS}(\psi) \to 0$$

in probability.

Proof See Appendix B.3. ■

Remark 3. *It should be noted that we have not used samples $x^{p^*} \sim p^*$ in the test statistic for the proposition above. If one were to use x^{p^*} in the test statistic, the proof strategy can be repeated up to the final sum of $A_n + B_n - 2C_n$, which instead will be*

$$\underbrace{\mathbb{E}\left[l(Y,Y')\right]\mathbb{E}\left[k(X,X')\right]}_{A_n} + \underbrace{\mathbb{E}\left[l(Y,Y')\right]\mathbb{E}\left[k(X^{p^*},X'^{p^*})\right]}_{B_n} - 2\underbrace{\mathbb{E}\left[l(Y,Y')\right]\mathbb{E}\left[k(X,X^{p^*})\right]}_{C_n}$$

which generally will not sum to 0 as $n \to \infty$, implying that we lose consistency. Thus, we use samples x^{p^*} to empirically improve power at the cost of being biased.

It is worth noting that the assumptions here are a couple of steps removed from the practical algorithm – in particular, they require access to true weights and a perfect permutation strategy. However, the results do indicate that we can expect the power of the test to improve with sample size when weights are estimated in a consistent manner and when the clustering approach to permutations consists of clusters with approximately equal conditional densities.

4. Estimation of Weights

Since we will generally never have access to the true weights needed for backdoor adjustment, we need to estimate them from observed data. We denote estimated weights as $\tilde{w}_i$. In this section, we describe how to estimate these $\tilde{w}_i$ for both categorical and continuous X.

4.1 Categorical treatment variable X

For categorical x_i we estimate w_i using the ratio $\frac{p^*(x_i)}{p(x_i|z_i)}$, where we take $p^*(x) = p(x)$. We first estimate $p(x)$ by simply taking the empirical probabilities for each category using the training data. For $p(x \mid z)$, we fit a probabilistic classifier mapping from z to each class of x. When X consists of multiple categorical dimensions, i.e. $d_X > 1$, we consider the $p^*(x)$ and $p(x \mid z)$ over the joint space of $\mathcal{X}_1 \times \ldots \times \mathcal{X}_{d_X}$. In the cases where d_X is large (≥ 8), we assume each x^d to be independent of each other and take $p(x) = \prod_{d=1}^{d_X} p(x_d)$ and $p(x \mid z) = \prod_{d=1}^{d_X} p(x_d \mid z)$.

4.2 Continuous treatment variable X

In the continuous case, we can no longer estimate $p(x \mid z)$ or $p^*(x)$ using a classifier straightforwardly. This complicates the estimation of the density ratio $w = \frac{p^*(x)}{p(x \mid z)}$, as we could

either estimate $p^*(x)$ and $p(x \mid z)$ separately using density estimation or estimate the density ratio directly. We review some existing methods below.

Direct density estimation of $p^*(x)$ and $p(x \mid z)$ allows for a broad range of methods such as normalizing flows (Rezende and Mohamed, 2015), generative adversarial networks (Goodfellow et al., 2014) and kernel density estimation (Botev et al., 2010) among many. While these methods provide accurate density estimation, they tend to be computationally expensive and hard to train (Mescheder et al., 2018). In this paper, we do not explore them further since densities themselves are not of direct interest.

RuLSIF (Yamada et al., 2011) propose using kernel ridge regression to directly estimate the density ratio $r(x) = \frac{p_1(x)}{p_2(x)}$ between distributions $p_1(x)$ and $p_2(x)$. While this method offers an analytical approach to estimating the density ratio, a regression may often not be flexible enough to learn our density ratio of interest in a high dimensional setting. We compare against RuLSIF in the experiment section.

Noise contrastive density estimation (NCE) (Gutmann and Hyvärinen, 2012) considers the problem of estimating an unknown density $p_{\text{true}}(\mathbf{x}; \theta)$ with parameters θ from samples $\mathbf{x} \in \mathbb{R}^d$. The key idea of NCE is to convert a density estimation problem to a classification problem by selecting an auxiliary noise contrastive distribution $p_{\text{noise}}(\mathbf{x})$ to compare with samples from p_{true}. This noise contrastive distribution is used to train a density ratio $\hat{p}(\mathbf{x}; \theta')$ with parametrization θ' of p_{true} to distinguish between fake samples $\mathbf{z} \sim p_{\text{noise}}$ and observations $\mathbf{x}_i \sim p_{\text{true}}$ through binary classification. We can then derive p_{true} by using this estimated density ratio. NCE has numerous desirable properties, including consistency under mild assumptions. We will propose and use a slightly modified NCE method to estimate our desired density ratio. We detail this method in the next section.

Telescoping density ratio (TRE) (Rhodes et al., 2020) considers the problem of estimating the density ratio between distributions p_0 and p_m using samples $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_m \sim p_m$. However, these density ratio problems tend to become pathological when p_0 and p_m are too far apart, exhibiting a phenomenon coined *density chasm*. The main idea of TRE is then to decompose this density ratio into several sub-tasks through a telescoping product

$$\frac{p_0(\mathbf{x})}{p_m(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \cdots \frac{p_{m-2}(\mathbf{x})}{p_{m-1}(\mathbf{x})} \frac{p_{m-1}(\mathbf{x})}{p_m(\mathbf{x})},$$

and estimate each individual ratio with separate estimators $r_k(\mathbf{x}; \theta_k) \approx p_k(\mathbf{x})/p_{k+1}(\mathbf{x})$ for $k = 0, \ldots, m-1$; we can then compose the original density ratio as

$$r(\mathbf{x}; \theta) = \prod_{k=0}^{m-1} r_k(\mathbf{x}; \theta_k) \approx \prod_{k=0}^{m-1} \frac{p_k(\mathbf{x})}{p_{k+1}(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_m(\mathbf{x})}.$$

To train each estimator $r_k(\mathbf{x}; \theta_k)$ we require a gradual transformation of samples between $\mathbf{x}_0$ and $\mathbf{x}_m$, resulting in intermediate samples $\mathbf{x}_k \sim p_k, \quad k = 0, \ldots, m-1$. We define these samples as a linear combination of $\mathbf{x}_0$ and $\mathbf{x}_m$

$$\mathbf{x}_k = \sqrt{1 - \alpha_k^2}\mathbf{x}_0 + \alpha_k \mathbf{x}_m, \quad k = 0, \ldots, m$$

where the α_k's form an increasing sequence from 0 to 1. The training objective

$$\mathcal{L}_{\mathrm{TRE}}(\theta) = \frac{1}{m}\sum_{k=0}^{m-1}\mathcal{L}_k\left(\theta_k\right)$$
$$\mathcal{L}_k\left(\theta_k\right) = -\mathbb{E}_{\mathbf{x}_k\sim p_k}\log\left(\frac{r_k\left(\mathbf{x}_k;\theta_k\right)}{1+r_k\left(\mathbf{x}_k;\theta_k\right)}\right) - \mathbb{E}_{\mathbf{x}_{k+1}\sim p_{k+1}}\log\left(\frac{1}{1+r_k\left(\mathbf{x}_{k+1};\theta_k\right)}\right)$$

is the average of all m losses of the subtasks.

4.3 NCE for bd-HSIC

In the do-null context, we have access to samples $\{(x_i, y_i, z_i)\}_{i=1}^n \sim p$. To calculate $\|\widehat{C}_{p^*}\|^2$ we need to estimate the density ratio $\tilde{w}_i = \frac{p^*(x_i)}{p(x_i|z_i)}$ from our observations. Here $p^*(x)$ is a chosen marginal distribution of X. We can express the density ratio as

$$w_i = \frac{p^*(x_i)}{p(x_i|z_i)} = \frac{p^*(x_i)p(z_i)}{p(x_i,z_i)}.$$

If we take $p^*(x) = p(x)$, the problem translates into finding the density ratios between the product of the marginals and the joint density.

We can make use of the NCE framework by taking joint samples $D_1 = \{(x_i, z_i)\}_{i=1}^{n_1}$ and approximate samples from the product of the marginals $D_2 = \left\{\left(x_i, z_{\pi(i)}\right)\right\}_{i=1}^{n_2} \sim p(x)p(z)$, for a randomly drawn permutation π. We obtain D_1 and D_2 by splitting the data set to ensure independence between positive samples and negative samples in NCE.

By setting $p_{\mathrm{true}}(x,z) = p(x,z)$ and $p_{\mathrm{noise}}(x,z) = p(x)p(z)$, we can parametrize the noise contrastive classifier as $h(x,z;\theta) = \sigma(\ln(\frac{1}{\nu}\frac{p(x)p(z)}{p(x,z;\theta)})) = \sigma(\ln(\frac{1}{\nu}r(x,z;\theta)))$, where $r(x,z;\theta)$ is a classifier parametrized by θ, $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\nu = \frac{n_2}{n_1}$, the ratio between samples from the product of the marginal to the joint distribution. By using NCE to directly estimate the density ratio between the product of the marginals and the joint density we gain the advantage that we do not need to specify an explicit noise contrastive distribution (and potentially introduce bias), as we can already obtain samples approximately through permutation. The problem is then reduced to a classification problem where we have to discriminate between samples from the product of marginals and the joint distribution. We introduce two modifications which take advantage of using the marginal $p^*(x)$.

4.3.1 NCE-p^*

Consider any $p^*(x) \neq p(x)$. We then build a classifier to discriminate data sets $D_1 = \{(x_i, z_i)\}$ from $D_2^{p^*} = \{(x_j^{p^*}, z_j)\}$ where $\{x_j^{p^*}\} \sim p^*$ independently of z_j, so $D_2^{p^*}$ contains samples from $p^*(x)\,p(z)$. When we pass any new pair (x,z) (i.e. regardless where it comes from, and in particular it can come from $p(x,z)$) to the classifier, it gives us the density ratio

$$\frac{p^*(x)p(z)}{p(x,z)} = \frac{p^*(x)}{p(x\,|\,z)},$$

which is then parametrized as

$$h(x,z;\theta) = \sigma\left(\ln\left\{\frac{1}{\nu}\frac{p^*(x)p(z)}{p(x,z;\theta)}\right\}\right) = \sigma\left(\ln\left\{\frac{1}{\nu}r(x,z;\theta)\right\}\right).$$

4.3.2 TRE-p^*

We can apply TRE to density ratio estimations between joint samples $\mathbf{x}_0 = (x, z)$ and product of marginals $\mathbf{x}_m = \left(X^{p^*}, z\right)$. For our particular context involving a chosen $p^*(x)$, we generate intermediate samples $\mathbf{x}_k = \left(\sqrt{1-\alpha_k^2}x + \alpha_k X^{p^*}, z\right)$ by fixing z for $k = 0, \ldots, m$.

It should be noted that neither TRE-p^* or NCE-p^* provides any guarantees in regard to uniform convergence rates of consistency.

4.4 Mixed treatment X

When X contains both continuous and categorical treatments, modifications to the continuous method are needed. We observe that if we take $X = x_{\text{cat}} \cup x_{\text{cont}}$ we have

$$w(x_{\text{cat}}, x_{\text{cont}}) = \frac{p(x_{\text{cat}}, x_{\text{cont}})}{p(x_{\text{cat}}, x_{\text{cont}} \mid z)} = \frac{p(x_{\text{cat}}, x_{\text{cont}})}{p(x_{\text{cat}} \mid z, x_{\text{cont}})p(x_{\text{cont}} \mid z)} \tag{5}$$

$$= \underbrace{\frac{p(x_{\text{cat}} \mid x_{\text{cont}})}{p(x_{\text{cat}} \mid z, x_{\text{cont}})}}_{\text{Classifiers}} \cdot \underbrace{\frac{p(x_{\text{cont}})}{p(x_{\text{cont}} \mid z)}}_{\text{NCE}}. \tag{6}$$

We note that we can decompose the density ratio into a *product* of density ratios estimated using classifiers and density ratio estimation methods. We will use this as the main method for mixed treatment data since this composition allows us to simplify the problem by avoiding estimating density ratios over joint categorical and continuous treatment data which could induce density chasms (Rhodes et al., 2020). We compare our proposed method to dimension-wise *mixing*, proposed in Rhodes et al. (2020). The same techniques can also be applied to NCE-p^*.

4.5 Algorithmic procedure

We describe the procedures for estimating weights $\tilde{w}$ and our proposed testing procedure.

4.5.1 Training the density ratio estimator

We train our density ratio estimators $r(x, z; \theta)$ through gradient descent. We parameterize all our estimators as Neural Networks (NN), due to their ability to fit almost any function. We summarize the training procedure for categorical data in Algorithm 2 and continuous data in Algorithm 3. We take our validation criteria in Algorithm 3 to be out-of-sample loss.

4.5.2 Testing procedure

The entire procedure of the test can be summarized in Algorithm 4.

It should be noted that we partition the data such that the data used for the estimation of weights are independent of the data used for the permutation test.

5. Simulations

We run experiments using bd-HSIC in the following contexts of do-null testing:

Algorithm 2: Training a density ratio estimator for categorical X

Input: Data $\mathcal{D} = \{x_i, z_i, x_i^{p^*}\}_{i=1}^{n//2}$
Output: Trained estimator $r(x, z; \theta')$
Partition data into training and validation $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{val}$
Initialise estimator $r(x, z, \theta')$
if $D > 8$ **then**
  for *each categorical* x^d **do**
    Estimate $p(x^d)$ using empirical probabilities
    Estimate $p(x^d \mid z)$ using a classifier with parameters θ_d
    Set $r^d(x^d, z; \theta_d) = \frac{p(x^d)}{p(x^d \mid z; \theta_d)}$
  end
  Set $r(x, z) = \prod_{d=1}^{n} r^d(x^d, z; \theta_d)$
end
else
  Estimate $p(x)$ over joint space
  Estimate $p(x \mid z)$ using a classifier θ
  Set $r(x, z) = \frac{p(x)}{p(x \mid z; \theta)}$
end
Return: Trained estimator $r(x, z)$

Algorithm 3: Training an NCE-based density ratio estimator

Input: Data $\mathcal{D} = \{x_i, z_i, x_i^{p^*}\}_{i=1}^{n//2}$
Output: Trained estimator $r(x, z; \theta)$
Partition data into training and validation $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{val}$
Initialize estimator $r(x, z, \theta)$
Partition data into joint and product of the margin samples $\mathcal{D}_{tr} = \mathcal{D}_{tr}^{\text{pom}} \cup \mathcal{D}_{tr}^{\text{joint}}$, $\mathcal{D}_{val} = \mathcal{D}_{val}^{\text{pom}} \cup \mathcal{D}_{val}^{\text{joint}}$
while *validation criteria* ν *not converged* **do**
  Sample positive and negative samples $\delta_+ \subset \mathcal{D}_{tr}^{\text{pom}}, \delta_- \subset \mathcal{D}_{tr}^{\text{joint}}$;
  Calculate loss $l = \mathcal{L}(r(\delta_+), r(\delta_-))$;
  Gradient Descent $\theta = \theta + \frac{\partial l}{\partial \theta}$;
  Calculate validation loss $\mathcal{L}(r(\mathcal{D}_{val}^{\text{pom}}), r(\mathcal{D}_{val}^{\text{joint}}))$;
end
Return: Trained estimator $r(x, z)$

1. Linear X, Y dependencies under multiple treatments and treatment types, multiple confounders and multiple outcomes

2. Non-linear X, Y dependencies under multiple treatments, multiple confounders and multiple outcomes

Algorithm 4: Testing $H_0 : p(y \mid do(x)) = p^*(y)$

Input: Data $\{(x_i, y_i, z_i)\}_{i=1}^n$, Distribution $p^*(x)$, density ratio estimator $r(\cdot)$, number of permutations n_q

Output: p-value for H_0

Find optimal c_{p^*} for continuous X

Sample $\{X_i^{p^*}\}_{i=1}^n \sim p^*$

Partition data into $\mathcal{D}_1 = \left\{(x_i, y_i, z_i, X_i^{p^*})\right\}_{i=1}^{\lfloor n/2 \rfloor}$ and $\mathcal{D}_2 = \left\{(x_i, y_i, z_i, X_i^{p^*})\right\}_{i=\lfloor n/2 \rfloor+1}^{n}$

Train r on $\mathcal{D}_1$

Estimate $p(x \mid z)$ on $\mathcal{D}_1$ using modified k-means

Get weights $\{\tilde{w}_i\}_{i=\lfloor n/2 \rfloor+1}^n = r(\left\{(x_i, z_i, X_i^{p^*})\right\}_{i=\lfloor n/2 \rfloor+1}^{n})$

Calculate $\|\widehat{C}_{p^*}\|^2$ using $\mathcal{D}_2$

Calculate permuted test statistics $\{\|\widehat{C}_{p^*}\|_1^2, \ldots, \|\widehat{C}_{p^*}\|_{n_q}^2\}$

Calculate the p-value as

$$p = 2 \min\left(1 - \frac{1+\sum_{i=1}^{n_q} 1_{\|\widehat{C}_{p^*}\|^2 < \|\widehat{C}_{p^*}\|_i^2}}{1+n_q}, \frac{1+\sum_{i=1}^{n_q} 1_{\|\widehat{C}_{p^*}\|^2 < \|\widehat{C}_{p^*}\|_i^2}}{1+n_q}\right)$$

Return: p

To extend the exposition of bd-HSIC, we contrast against post-double selection (PDS), which serves as a representative benchmark with pathologies that bd-HSIC attempts to amend.

5.1 Comparison against semi-parametric methods

We compare against the popular post-double selection (PDS) method (Belloni et al., 2014), a semi-parametric lasso-based model that is widely used for causal estimation. PDS considers the following setup

$$\begin{aligned} y_i &= d_i \alpha_0 + g(z_i) + \zeta_i, & \mathbb{E}[\zeta_i \mid z_i, d_i] = 0 \\ d_i &= m(z_i) + v_i, & \mathbb{E}[v_i \mid z_i] = 0 \end{aligned} \tag{7}$$

where y_i is the outcome variable, d_i is the policy/treatment variable whose impact α_0 is the quantity of interest, z_i represents confounding factors , and ζ_i and v_i are disturbances. The problem is then recast into a linear form

$$\begin{aligned} y_i &= d_i \alpha_0 + \underbrace{x_i' \beta_{g0} + r_{gi}}_{g(z_i)} + \zeta_i, \\ d_i &= \underbrace{x_i' \beta_{m0} + r_{mi}}_{m(z_i)} + v_i, \end{aligned} \tag{8}$$

where $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ are approximations to $g(z_i)$ and $m(z_i)$, and r_{gi} and r_{mi} are the corresponding approximation errors. PDS then uses lasso to estimate $m(z_i)$ and $g(z_i)$ and for selecting non-zero control variables, and then regresses y_i on d_i together with the union of

selected non-zero control variables. Note that PDS is a lasso-based model, which makes it susceptible to non-linear confounding effects and causal effects. Given the above contexts, the coverage of PDS includes linear X, Y dependencies with multiple treatments under multiple confounders.

To make comparisons straightforward, we only compare to PDS in univariate treatment, confounder and outcome cases.

5.2 Comparisons against CfME and Singh et al. (2023)

We additionally compare against CfME proposed in Muandet et al. (2021) for the binary treatment case and Singh et al. (2023) for general treatment. While it is not immediately clear how to adapt Singh et al. (2023) for a hypothesis test, we provide a derivation in Appendix E. We will refer to the newly derived test as *backdoor-CME* (bd-CME). For both these cases, we consider direct permutation on Y following the implementation of CfME.

An inherent limitation of CfME is that it only can be used for binary treatment cases. For bd-CME, which is entirely kernel dependent, problems could arise when the data consists of both categorical and continuous variables, as it becomes unclear how to adequately select kernels for both of these data types. One could of course consider a product kernel, but such a kernel may break certain dependencies that render the test invalid.

5.2.1 Ablation study

We further compare to $\tilde{w}_i \sim \text{Uniform}(0, 1)$. While this choice of weights is not a very principled approach, it serves as a reference to see whether bd-HSIC actually needs correctly estimated weights to have power.

5.2.2 Studying the size under H_0

When the null hypothesis is true, we would expect the p-value distribution of the test run on several data sets to have the correct size. Here we use level $\alpha = 0.05$.

5.2.3 Studying the power under H_1

Under the alternative, a desired property is high power across all parameters of the data generation. We will conduct experiments to demonstrate when the test has high power and when it does not. We calculate power for level $\alpha = 0.05$.

We present our results in bivariate plots, where we plot β_{XY} (x-axis) against the power at level $\alpha = 0.05$ (y-axis). An ideal plot would be a discontinuous function with height 0.05 at $\beta_{XY} = 0$ and height 1.0 for $\beta_{XY} > 0$. We would then expect that as the sample size increases, the plot would approach this ideal.

5.3 Results

We divide our experiments into two steps: we first take the *true weights* and use them in the subsequent permutation test for bd-HSIC. In this initial step, we chose $p^* = p$. This is

intended as a unit test to validate that the parameter selection used for data generation is working. It should be noted that generally the choice of p^* and estimation of c_{p^*} must be done cautiously to avoid double use of data.

In the second step we estimate weights from the data and investigate the effectiveness of our proposed density ratio estimation procedure. We consider the following methods for weight estimation: RuLSIF, random uniform, NCE-p^* and TRE-p^* as described in Section 4.3 and Section 4.

5.3.1 Binary treatment X

We simulate univariate binary data according to Appendix F.1. We vary the dependence β_{XY} between X and Y to be $\{0.0, 0.005, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1\}$. We plot the power for the level $\alpha = 0.05$ against β_{XY} in Figure 5.

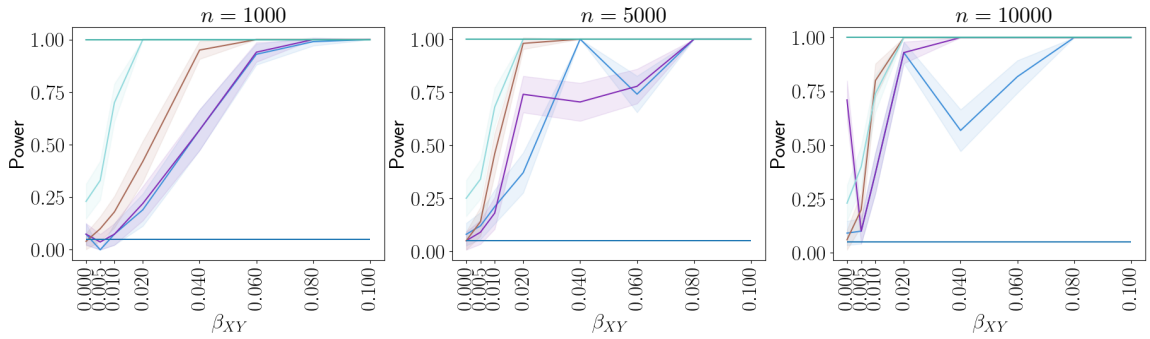


Figure 5. bd-HSIC(NCE-p^*): bd-HSIC(true weights): PDS: CfME: bd-CME:
Binary treatment for $n = 1000, 5000, 10000$ using an RBF kernel in bd-HSIC. Both CfME and bd-CME has incorrect size. The horizontal line denotes when power is 0.05 as a visual reference.

5.3.2 Continuous treatment X

Linear dependency between X and Y
The data are simulated for $(d_z, d_x, d_y) \in \{(1,1,1), (3,3,3), (15,3,3), (50,3,3)\}$. In our experiments, we found that a strong confounding effect (i.e. large β_{XZ}) led to a smaller effective sample size, making it harder to obtain a consistent test under H_0. For H_1, the difficulty was mostly controlled by the magnitude of β_{XY}, where small magnitudes often led to a test with little or no power. We have chosen rejection sampling parameters $\theta, \phi, \beta_{XZ}, \beta_{YZ}$ such that the tests are non-trivial but not a failure mode, where θ, ϕ control variance of the marginal distribution of the treatment and the variance of proposal distribution respectively. For exact simulation details, we refer to the appendix. We consider $\beta_{XY} \in [0.0, 0.05]$, with $\beta_{XY} = 0.0$ corresponding to H_0. We present results for continuous treatment in Figure 6. We note that bd-CME has inflated type 1 errors for $d_Z = 15$.
Non-linear dependency between X and Y
Here we simulate data for $(d_z, d_x, d_y) \in \{(1,1,1), (50,3,3)\}$ using the same parameters as in

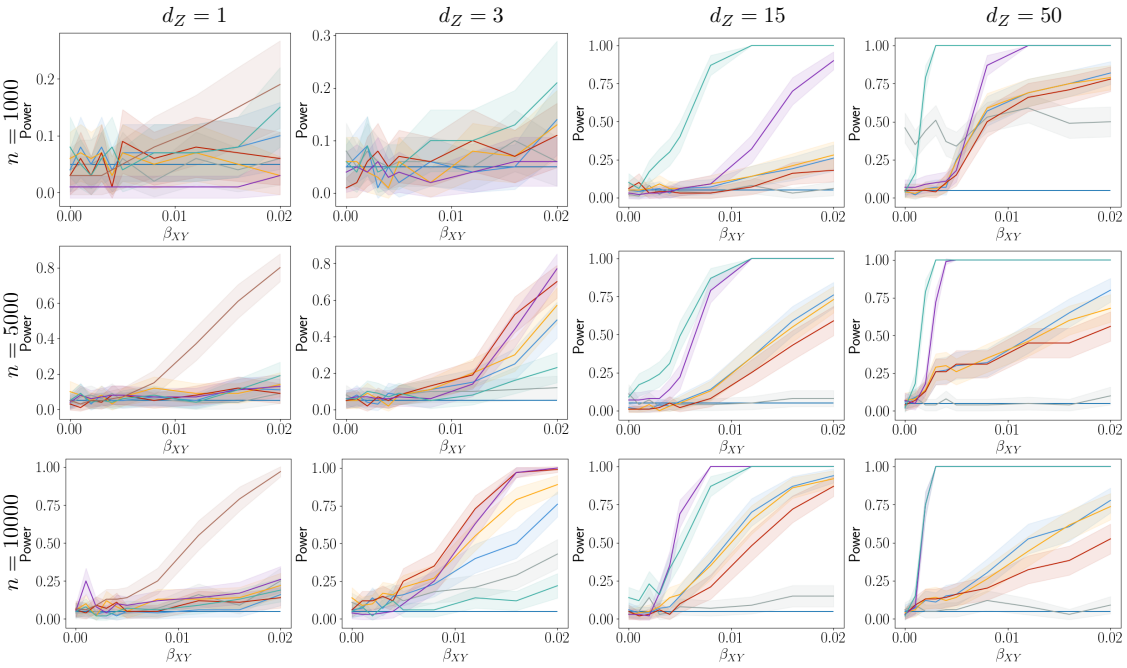
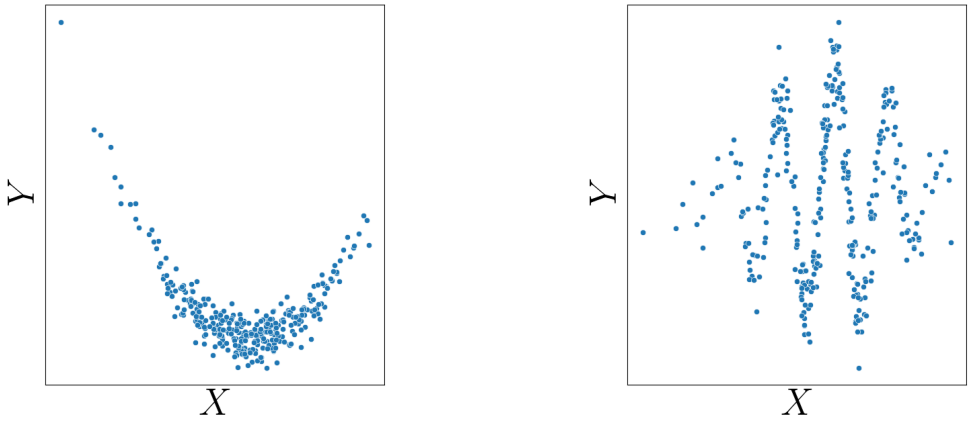


FIGURE 6. bd-HSIC(NCE-p^*): bd-HSIC(TRE-p^*): bd-HSIC(random uniform): bd-HSIC(RuLSIF): bd-HSIC(true weights): PDS: bd-CME:
Continuous treatment results. We find that RuLSIF has incorrect size under the null and that uniform weights has less power. NCE-p^* seems to have the best power while being calibrated under the null. We generally note that random uniform weights have less or no power when compared to "true weights" and estimated weights.



(A) U-shaped dependency, $Y = X^2\beta_{XY}$

(B) General symmetric non-linear dependency, $Y = \exp(-0.1X^2)\cos(X\pi)\beta_{XY}$

FIGURE 7. Samples from p^* under H_1 where the dependency is non-linear.

the linear case. The only difference now is that we consider a non-linear dependency between X and Y illustrated in Figure 7. Figure 7a illustrates a U-shaped dependency between X and Y, which can be found in relationships between happiness vs. age (Kostyshak, 2017), and BMI vs. fragility (Watanabe et al., 2020) to name a few. The U-shaped dependency can be generalized to symmetric non-linear relationships between X and Y illustrated in Figure 7b. We show the results in Figure 8 and Figure 9. We note that PDS has no power against the alternative when the dependency between X and Y is symmetric and non-linear, which is expected due to the linear nature of PDS.

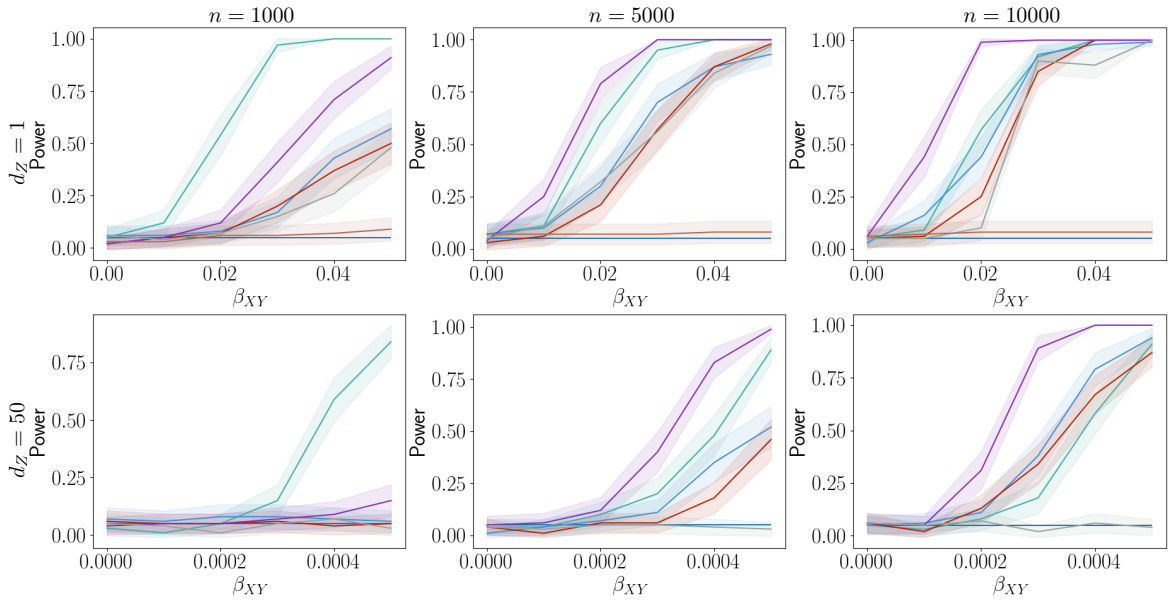


Figure 8. bd-HSIC(NCE-p^*): bd-HSIC(random uniform): bd-HSIC(RuLSIF): bd-HSIC(true weights): PDS: bd-CME:

Experiments for U-shaped dependency between X and Y.

5.3.3 Mixed treatment X

The data are simulated for $(d_z, d_x, d_y) \in \{(2,2,2), (4,4,3), (15,6,6), (50,8,8)\}$. We simulate the mixed data according to Algorithm F.3. Here we fix half of the X's to be continuous and the other half binary. We consider $\beta_{XY} \in [0.0, 0.1]$. We follow the same principles as in the continuous case when selecting $\theta, \phi, \beta_{XZ}, \beta_{YZ}$.

In our experiments, we compare against RuLSIF and randomly sampled uniform weights. We also compare between estimating the density ratio of the binary treatments separately (denoted with suffix "prod") and all treatments simultaneously (no suffix). The "mixed" suffix is a reference to the *dimension-wise mixing* proposed in Rhodes et al. (2020), which is applied when using TRE-p^*. We present the results in Figure 10. We note that bd-CME has an inflated type 1 error rate for $d_Z = 50$.

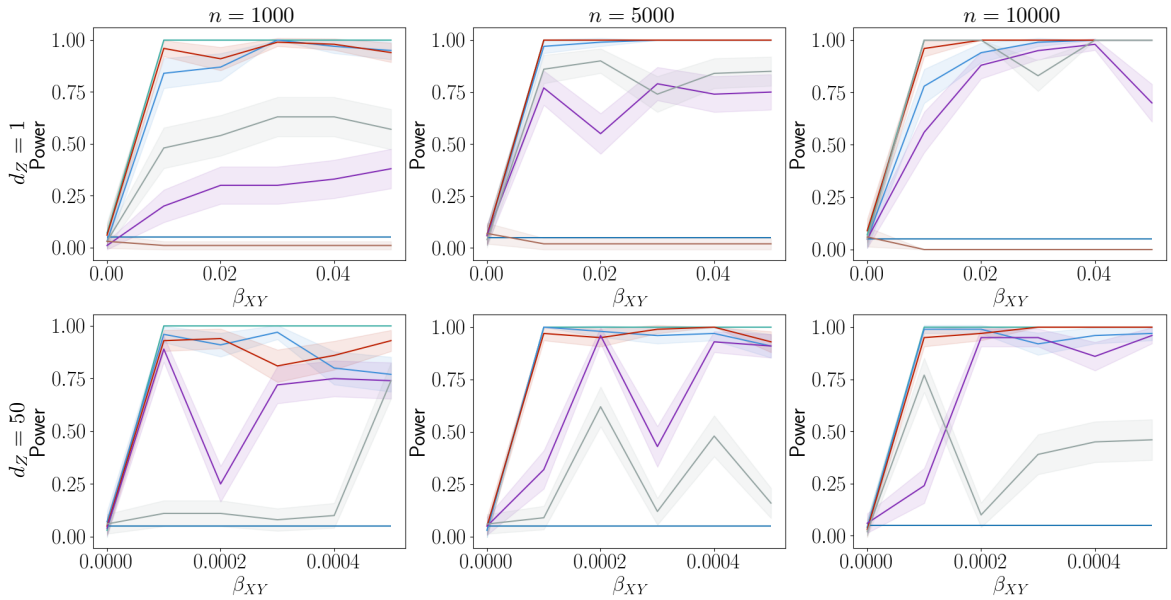


FIGURE 9. bd-HSIC(NCE-p^*): bd-HSIC(random uniform): bd-HSIC(RuLSIF): bd-HSIC(true weights): PDS: bd-CME:

Experiments for general non-linear symmetric dependency between X and Y.

5.4 Pitfalls

5.4.1 Choice of kernels, a cautionary tale

To improve the power of bd-HSIC, we could instead use a linear kernel. Figure 11 illustrates that bd-HSIC has comparable power to PDS, when using a linear kernel for testing the do-null when one considers the univariate binary treatment case.

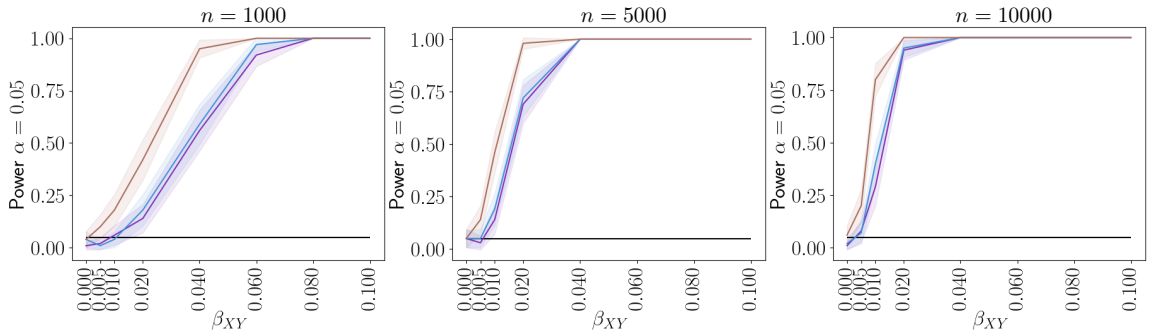


FIGURE 11. bd-HSIC(NCE-p^*): bd-HSIC(true weights): PDS:
Binary treatment for $n = 1000, 5000, 10000$ using an linear kernel in bd-HSIC. Compared to Figure 5, bd-HSIC now has similar power to PDS.

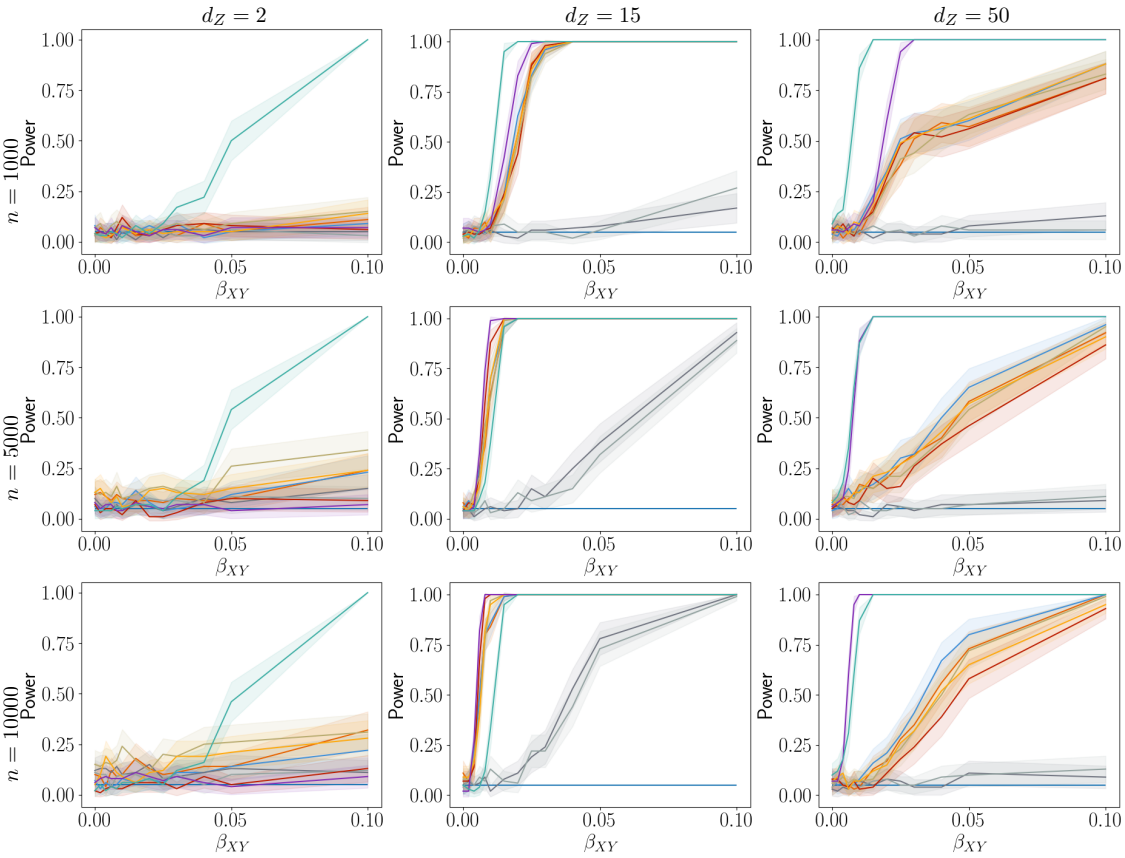
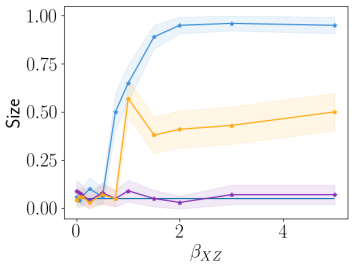


FIGURE 10. bd-HSIC(NCE-p^*, mixing): bd-HSIC(TRE-p^*, mixing): bd-HSIC(random uniform): bd-HSIC(RuLSIF): bd-HSIC(true weights): bd-HSIC(NCE-p^* prod): bd-HSIC(TRE-p^* product): bd-HSIC(RuLSIF product): bd-CME:
Mixed treatment results. We find that RuLSIF has an incorrect size under the null. TRE-p^* prod seems to have the best power while being calibrated under the null.

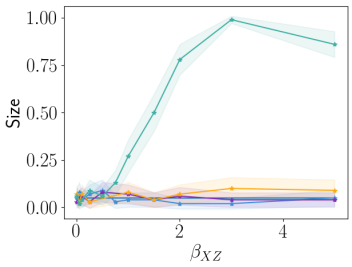
However, when applying the linear kernel to univariate data for a continuous treatment the test becomes uncalibrated. In fact, the linear kernel makes bd-HSIC much more sensitive to confounding, exhibited in Figure 12a.

5.4.2 When does bd-HSIC break?

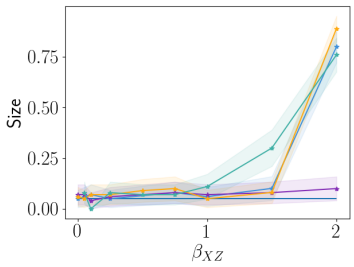
We demonstrate a typical failure mode of bd-HSIC, when the value of β_{XZ} is so strong that the density ratio estimation fails. We show that NCE-p^* and TRE-p^* have incorrect size when β_{XZ} becomes large enough in Figure 12c. Here we generate data under the null for $\beta_{XZ} \in \{0.0, 0.05, 0.1, 0.15, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$ for $d_X = d_Y = 3, d_Z = 15$.



(a) Applying a linear kernel to univariate continuous treatment data. The linear kernel may cause the test to have incorrect size.



(b) Applying a RBF kernel to univariate continuous treatment data.



(c) bd-HSIC with estimated density ratios exhibits incorrect size for a certain amount of confounding. True weights are consistent.

Figure 12. bd-HSIC(NCE-p^*): bd-HSIC(TRE-p^*): bd-HSIC(true weights): bd-CME:

5.5 Experiments on real-world data

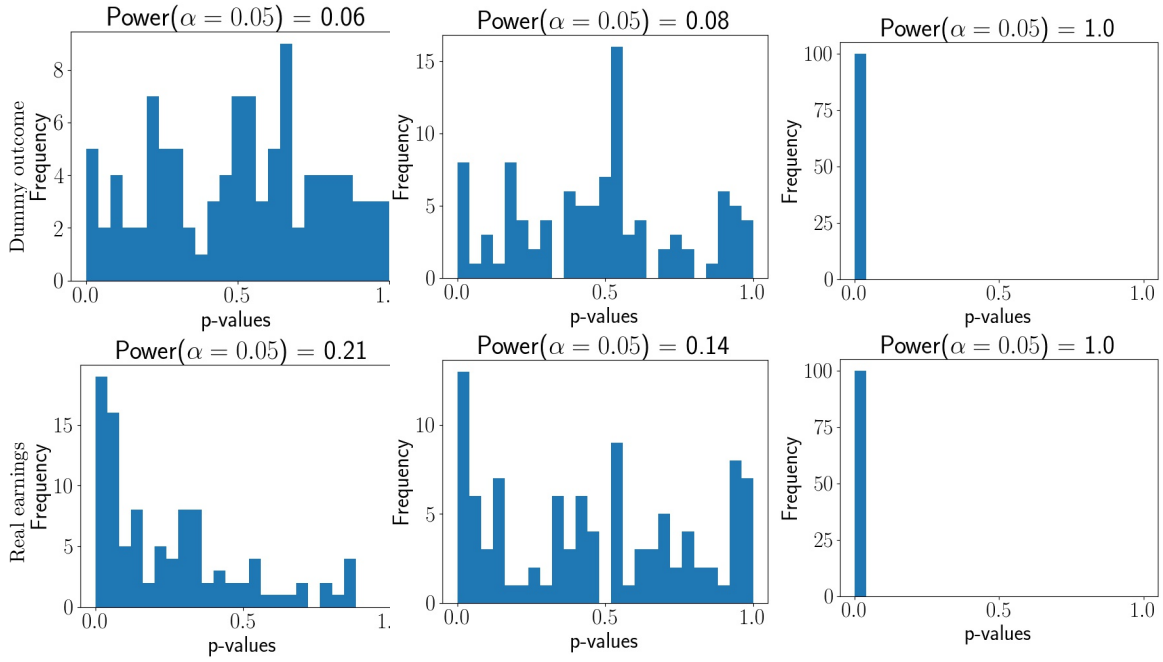
We apply bd-HSIC to two real-world data and compare against PDS and bd-CME.

5.5.1 Lalonde data set experiments

The Lalonde data set comes from a study that looked at the effectiveness of a job training program (the treatment) on the real earnings of an individual, a couple of years after completion of the program (the outcome). Each individual has several descriptive covariates such as age, academic background, which confound the relationship between the treatment and the outcome. We compare the power between PDS and bd-HSIC on the Lalonde data set in Figure 13. This is done by calculating the p-value on 100 bootstrap sampled subsets of the data set. We further generate a random independent dummy outcome to verify that our tests are calibrated. We note that both PDS and bd-HSIC have the correct type 1 control for dummy outcome, while their power is comparable for the real earnings outcome.

5.5.2 Twins data set experiments

The twins data set (Louizos et al., 2017) considers data of twin births in the US between 1989–1991. Here the treatment is being born the heavier twin and the outcome is mor-



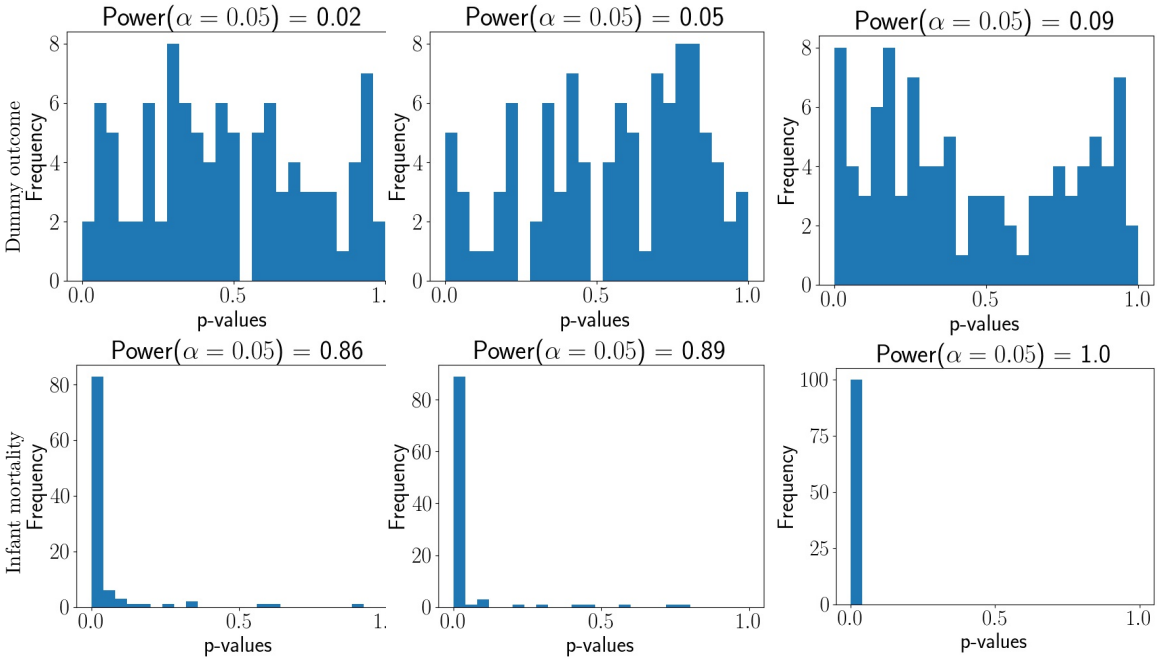
(A) PDS (B) bd-HSIC (TRE-p^*) (C) bd-CME

FIGURE 13. PDS suggests that there exists a causal association between receiving training and real earnings. bd-CME appears to have incorrect size.

tality. Besides treatment and outcome, there are also descriptive confounders such as the smoking habits of the parents, education level of the parents, and medical risk factors of the children among many. For our experiments, we construct a slight variation of the experiment presented in Louizos et al. (2017), where we instead take the treatment to be $T = \left(\text{Weight}_{\text{heavier twin}}, \text{Weight}_{\text{lighter twin}}, \text{Weight}_{\text{heavier twin}} - \text{Weight}_{\text{lighter twin}}\right)$ and the outcome to be in the set $\{-1, 0, 1\}$, where -1 indicates that the lighter twin died, 1 that the heavier twin died, and 0 that either, neither or both died; everything else is kept the same. Similar to the Lalonde data sets we calculate the p-value on 100 bootstrap sampled subsets of the data set and present results in Figure 14. Similar to the Lalonde data set, we generated a random independent dummy outcome to verify that our tests are calibrated. All three methods suggest there exists a causal association between infant weight and mortality.

6. Conclusion

We present a novel non-parametric method termed backdoor-HSIC (bd-HSIC), which is an importance-weighted covariance-based statistic to test the causal null hypothesis, or *do-null*. We first show that our proposed estimator for bd-HSIC is consistent. Experiments on a variety of synthetic data sets, including linear and non-linear dependencies, with different numbers of confounders and treatments, show that bd-HSIC is a flexible method with wider coverage of scenarios than parametric methods such as PDS. Finally, we compare bd-HSIC to PDS on two real-world data sets. Assuming a valid choice of confounders Z satisfying



(A) PDS (B) bd-HSIC (TRE-p^*) (C) bd-CME

FIGURE 14. All three methods suggest that there exists a causal association between weight and mortality.

the back door criterion, the test evaluates whether there is any effect on the distribution of the outcome.

A major benefit of bd-HSIC is that can serve as a powerful tool in causal inference as it complements parametric methods such as PDS. For example, PDS generally has better power when the underlying dependency is linear but fails if the dependency is symmetrically non-linear. Even Double Machine Learning, proposed in Chernozhukov et al. (2018) and a generalization of PDS, will not straightforwardly capture such non-linear causal dependencies without further assumptions. This is inherent to the design of Double Machine Learning, which is based on semi-parametric partially linear models. bd-HSIC is fully non-parametric and can be used as a general test for non-linear causal association, which is of broad interest to all statistically inclined sciences.

By combining with kernel conditional independence tests, this work could be extended to testing for null *conditional* average treatment effects, as well as to testing more general *nested* independence constraints (Richardson et al., 2023).

Acknowledgments

The authors would like to sincerely thank the reviewers and Silvia Chiappa for their feedback and suggestions that helped improve the paper. The authors are also thankful to Professor Tom Rainforth (University of Oxford) and Professor Ricardo Silva (University College London) for their suggestions. Robert Hu wrote the majority of the paper during his time as a PhD student at the Department of Statistics at the University of Oxford and was funded by Hennes & Mauritz AB during his studies.

Appendix A. Remark on $p^*(y)$

Why do we have to consider $H_0 : p(y|do(x)) = p^*(y)$ instead of $H_0 : p(y|do(x)) = p(y)$? We show in the proposition below that we need arbitrary distributions $p^*(y)$ to formulate the hypothesis in a distributional sense since we may have that the do-null hypothesis holds but $p(y \mid do(x)) \neq p(y)$.

Proposition 6. $p(y|do(x)) = p(y) \quad \not\Longleftarrow \quad Y \perp X \mid do(X = x)$.

Proof *We give an example as proof. Consider a backdoor setting. Let X, Y, Z all be binary random variables, and assume that X, Z are Bernoulli distributed with $p = 0.5$. Then let's consider the following distribution:*

$$\begin{aligned} p(Y = 1 \mid X = 1, Z = 1) &= p + \varepsilon \\ p(Y = 1 \mid X = 0, Z = 1) &= p \\ p(Y = 1 \mid X = 1, Z = 0) &= p - \varepsilon \\ p(Y = 1 \mid X = 0, Z = 0) &= p \end{aligned} \tag{9}$$

Calculating $p(Y = 1 \mid do(X = x)) = \sum_Z p(Y = 1 \mid X, Z)p(Z) = p$ for both $X = 0, 1$. Now let $p(X = 1 \mid Z) = \frac{1}{4} + \frac{Z}{2}$. But then the marginal

$$\begin{aligned} p(Y = 1) &= \sum_{X,Z} p(Y = 1 \mid X, Z)p(X \mid Z)p(Z) \\ &= 0.5(0.75(p + \varepsilon) + 0.25p + 0.25(p - \varepsilon) + 0.75p) = p + \frac{\varepsilon}{4} \end{aligned} \tag{10}$$

Thus we may have that the do-null hypothesis holds but $p(y \mid do(x)) \neq p(y)$.

■

Appendix B. Consistency proofs

B.1 Proof of Theorem 1

Proof We first define the kernel mean embeddings used in our proposed estimator:

$$\begin{aligned} \mathbb{E}[W_{p^*} k(x, \cdot) l(y, \cdot)] = \mu_{xy}(\cdot) &= \int W_{p^*} k(x, \cdot) l(y, \cdot) \, d\mathbb{P}_{\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{X}}(x, y, z) \\ &= \int \frac{p^*(x)}{p(x \mid z)} k(x, \cdot) l(y, \cdot) \, d\mathbb{P}_{\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{X}}(x, y, z) \end{aligned}$$

$$\mathbb{E}[k(X^{p^*},\cdot)] = \mu_{X^{p^*}}(\cdot) = \int k(X^{p^*},\cdot)\, d\mathbb{P}^*(x)$$

$$\mathbb{E}[W_{p^*} l(y,\cdot)] = \mu_y(\cdot) = \int \frac{p^*(x)}{p(x \mid z)} l(y,\cdot) d\mathbb{P}_{\mathcal{X}\otimes\mathcal{Y}\otimes\mathcal{X}}(x,y,z).$$

Then we write the following for shorthand:

$$C_{p^*} = \mu_{xy}(\cdot) - \mu_{X^{p^*}}(\cdot) \otimes \mu_y(\cdot)$$

Thus

$$\mathbb{E}\left[\|C_{p^*} - \widehat{C}_{p^*}\|^2_{\mathrm{HS}}\right]$$

Which is

$$\mathbb{E}\Bigg[\Bigg\|\mu_{xy}(\cdot) - \mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot) - \Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^n \frac{p^*(x_i)}{p(x_i \mid z_i)} k(\cdot,x_i)\otimes l(\cdot,y_i)}_{\hat{\mu}_{xy}} - \underbrace{\left(\frac{1}{n}\sum_{j=1}^n k\left(\cdot, x_j^{p^*}\right)\right)}_{\hat{\mu}_{X^{p^*}}} \otimes \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \frac{p^*(x_i)}{p(x_i \mid z_i)} l\left(\cdot, y_i\right)\right)}_{\hat{\mu}_y}\Bigg)\Bigg\|^2\Bigg].$$

Note that $w_i = \frac{p^*(x_i)}{p(x_i|z_i)}$, and for now we consider we have access to the true weights. Rearranging terms we get

$$\mathbb{E}\left[\left\|\underbrace{\mu_{xy}(\cdot) - \hat{\mu}_{xy}(\cdot)}_{A} + \underbrace{(\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot) - \mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot))}_{B}\right\|^2\right] = \mathbb{E}\left[\langle A,A\rangle + 2\langle A,B\rangle + \langle B,B\rangle\right].$$

The proof strategy is to obtain convergence rates for each term. First term $\langle A, A\rangle$:

$$\mathbb{E}\left[\langle A,A\rangle\right] = \mathbb{E}\left[\langle \mu_{xy}(\cdot), \mu_{xy}(\cdot)\rangle\right] - 2\mathbb{E}\left[\langle \mu_{xy}(\cdot), \hat{\mu}_{xy}(\cdot)\rangle\right] + \mathbb{E}\left[\langle \hat{\mu}_{xy}(\cdot), \hat{\mu}_{xy}(\cdot)\rangle\right]$$

Let $x', y', z' \sim p$ be an independent copy of $x, y, z \sim p$. The first term is then

$$\mathbb{E}\left[\langle \mu_{xy}(\cdot), \mu_{x'y'}(\cdot)\rangle\right] = \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)}\frac{p^*(x')}{p(x' \mid z')} k(x,x') l(y,y')\right],$$

the second one:

$$\begin{aligned}
\mathbb{E}\left[\langle \mu_{xy}(\cdot), \hat{\mu}_{xy}(\cdot)\rangle\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{p^*(x_i)}{p(x_i \mid z_i)} \mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x \mid z)} k(x,x_i) l(y,y_i)\right]\right] \\
&= \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x' \mid z')} \mathbb{E}_{x,y,z}[\frac{p^*(x)}{p(x \mid z)} k(x,x') l(y,y')]\right] \\
&= \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)}\frac{p^*(x')}{p(x' \mid z')} k(x,x') l(y,y')\right],
\end{aligned}$$

and the final one:

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j} w_i w_j k(x_i,x_j) l(y_i,y_j)\right] =$$
$$\underbrace{\mathbb{E}\left[\frac{1}{n^2}\sum_{i\neq j} w_i w_j k(x_i,x_j) l(y_i,y_j)\right]}_{(a)} + \underbrace{\mathbb{E}\left[\frac{1}{n^2}\sum_{i} w_i^2 k(x_i,x_i) l(y_i,y_i)\right]}_{(b)}.$$

Then (a) is

$$(a) = \frac{(n-1)}{n}\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\mid z)}\frac{p^*(x')}{p(x'\mid z')}k(x,x')l(y,y')\right].$$

(b):

$$\begin{aligned}(b) &= \frac{1}{n}\mathbb{E}\left[\left(\frac{p^*(x)}{p(x\mid z)}\right)^2 k(x,x)l(y,y)\right]\\ &\leq \frac{1}{n}\underbrace{\sup_{x\in X} k(x,x)\sup_{y\in Y} k(y,y)}_{\text{Assumed to be bounded by } C}\quad\underbrace{\mathbb{E}\left[\left(\frac{p^*(x)}{p(x\mid z)}\right)^2\right]}_{\text{finite variance of density ratio, i.e. bounded by some constant } D}\\ &= \frac{1}{n}CD\end{aligned}$$

So for $\langle A, A\rangle$ the convergence rate is $\mathcal{O}(\frac{1}{n})$. $\langle A, B\rangle$:

$$\begin{aligned}\langle A,B\rangle = &\underbrace{\langle \mu_{xy}(\cdot), \hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(1)} + \underbrace{\langle \hat{\mu}_{xy}(\cdot), \mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(2)}\\ &- \underbrace{\langle \hat{\mu}_{xy}(\cdot), \hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(3)} - \underbrace{\langle \mu_{xy}(\cdot), \mu_{X^{p^*}}(\cdot)\otimes\mu_{y'}(\cdot)\rangle}_{(4)}\end{aligned}$$

Thus

$$\begin{aligned}(1) &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\frac{p^*(x_j)}{p(x_j\mid z_j)}\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}k(x,x_i^{p^*})l(y,y_j)\right]\right]\\ &= \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}k(x,x_i^{p^*})l(y,y_j)\right]\right]\end{aligned}$$

$$\begin{aligned}(2) &= \mathbb{E}\left[\frac{1}{n}\sum_{i}^{n}\frac{p^*(x_i)}{p(x_i\mid z_i)}\mathbb{E}_{x_q}[k(x_i,X^{p^*})]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}l(y,y_i)\right]\right]\\ &= \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\mathbb{E}_{x_q}[k(x',X^{p^*})]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}l(y,y')\right]\right]\end{aligned}$$

$$
\begin{aligned}
(3) &= \mathbb{E}\left[\frac{1}{n^3}\sum_{i,k,j}\frac{p^*(x_i)}{p(x_i\mid z_i)}\frac{p^*(x_j)}{p(x_j\mid z_j)}k(x_i,X_k^{p^*})l(y_i,y_j)\right]\\
&= \mathbb{E}\left[\frac{1}{n^3}\sum_{k,i\neq j}\frac{p^*(x_i)}{p(x_i\mid z_i)}\frac{p^*(x_j)}{p(x_j\mid z_j)}k(x_i,X_k^{p^*})l(y_i,y_j)\right]\\
&\quad+\mathbb{E}\left[\frac{1}{n^3}\sum_{k,i=j}\left(\frac{p^*(x_i)}{p(x_i\mid z_i)}\right)^2k(x_i,X_k^{p^*})l(y_i,y_i)\right]\\
&= \frac{n-1}{n}\mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}\mathbb{E}_{X^{p^*}}\left[k(x,X^{p^*})\right]l(y,y')\right]\right]\\
&\quad+\frac{1}{n}\mathbb{E}_{X^{p^*},x,y,z}\left[\left(\frac{p^*(x)}{p(x\mid z)}\right)^2k(x,X^{p^*})l(y,y)\right].
\end{aligned}
$$

$$
(4) = \mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}\mathbb{E}_{x_q}[k(x,X^{p^*})]\mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}l(y,y')\right]\right]
$$

We can use the same argument as in $\langle A,A\rangle$, consequently $\langle A,B\rangle \propto \mathcal{O}(\frac{1}{n})$ Let $X^{p^*\prime},x',y',z'$ be independent copies of X^{p^*},x,y,z. Then

$$
\begin{aligned}
\langle B,B\rangle = \underbrace{\langle\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot),\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(1)} &- \underbrace{2\langle\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot),\mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(2)}\\
&+\underbrace{\langle\mu_{X^{p^*\prime}}(\cdot)\otimes\mu_{y'}(\cdot),\mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(3)}
\end{aligned}
$$

Where

$$
\begin{aligned}
&(1)\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{u,v} k(X_u^{p^*}, X_v^{p^*})\frac{1}{n^2}\sum_{i,j}\frac{p^*(x_i)}{p(x_i \mid z_i)}\frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j)\right]\\
&= \mathbb{E}\Bigg[\frac{1}{n^2}\left(\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*}) + \sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\right)\\
&\quad\times\frac{1}{n^2}\left(\sum_{i\neq j}\frac{p^*(x_i)}{p(x_i \mid z_i)}\frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j) + \sum_{i=j}\left(\frac{p^*(x_i)}{p(x_i \mid z_i)}\right)^2 l(y_i, y_i)\right)\Bigg]\\
&= \underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i\neq j}\frac{p^*(x_i)}{p(x_i \mid z_i)}\frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j)\right]}_{a}\\
&\quad+\underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i=j}\left(\frac{p^*(x_i)}{p(x_i \mid z_i)}\right)^2 l(y_i, y_i)\right]}_{b}\\
&\quad+\underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i=j}\left(\frac{p^*(x_i)}{p(x_i \mid z_i)}\right)^2 l(y_i, y_i)\right]}_{c}\\
&\quad+\underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i\neq j}\frac{p^*(x_i)}{p(x_i \mid z_i)}\frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j)\right]}_{d}.
\end{aligned}
$$

Each term is then:

$$
\begin{aligned}
a &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*}) \sum_{i\neq j} \frac{p^*(x_i)}{p(x_i \mid z_i)} \frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j)\right] \\
&= \frac{(n-1)^2}{n^2} \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x' \mid z')} \frac{p^*(x)}{p(x \mid z)} l(y, y')\right] \\
b &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*}) \sum_{i=j} \left(\frac{p^*(x_i)}{p(x_i \mid z_i)}\right)^2 l(y_i, y_i)\right] \\
&= \frac{1}{n^2} \mathbb{E}_{X^{p^*}}\left[k(X^{p^*}, X^{p^*})\right] \mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x \mid z)}\right)^2 l(y, y)\right] \\
c &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*}) \sum_{i=j} \left(\frac{p^*(x_i)}{p(x_i \mid z_i)}\right)^2 l(y_i, y_i)\right] \\
&= \frac{n-1}{n^2} \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x \mid z)}\right)^2 l(y, y)\right] \\
d &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*}) \sum_{i\neq j} \frac{p^*(x_i)}{p(x_i \mid z_i)} \frac{p^*(x_j)}{p(x_j \mid z_j)} l(y_i, y_j)\right] \\
&= \frac{n-1}{n^2} \mathbb{E}_{X^{p^*}}\left[k(X^{p^*}, X^{p^*})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x' \mid z')} \frac{p^*(x)}{p(x \mid z)} l(y, y')\right]
\end{aligned}
$$

$$
\begin{aligned}
(2) &= \mathbb{E}\left[\frac{1}{n}\sum_i \mathbb{E}_{X^{p^*}}[k(X^{p^*}, X_i^{p^*})] \frac{1}{n}\sum_j \frac{p^*(x_j)}{p(x_j \mid z_j)} \mathbb{E}_{x,y,z}[\frac{p^*(x)}{p(x \mid z)} l(y, y_j)]\right] \\
&= \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)} \frac{p^*(x')}{p(x' \mid z')} l(y, y')\right]
\end{aligned}
$$

$$
(3) = \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)} \frac{p^*(x')}{p(x' \mid z')} l(y, y')\right]
$$

We note that the $\mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)} \frac{p^*(x')}{p(x' \mid z')} l(y, y')\right]$-terms collapse for $(1), (2), (3)$. Thus:

$$
\begin{aligned}
\langle B, B\rangle &= \left(\frac{-2}{n} + \frac{1}{n^2}\right) \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x \mid z)} \frac{p^*(x')}{p(x' \mid z')} l(y, y')\right] \\
&+ \frac{1}{n^2} \mathbb{E}_{X^{p^*}}\left[k(X^{p^*}, X^{p^*})\right] \mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x \mid z)}\right)^2 l(y, y)\right] \\
&+ \frac{n-1}{n^2} \mathbb{E}_{X^{p^*}, X^{p^{*\prime}}}\left[k(X^{p^*}, X^{p^{*\prime}})\right] \mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x \mid z)}\right)^2 l(y, y)\right] \\
&+ \frac{n-1}{n^2} \mathbb{E}_{X^{p^*}}\left[k(X^{p^*}, X^{p^*})\right] \mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x' \mid z')} \frac{p^*(x)}{p(x \mid z)} l(y, y')\right]
\end{aligned}
$$

It suffices now to upper bound all the remaining expectations. First note that $\mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right] \leq \sup_{X^{p^*},X^{p^{*\prime}}} k(X^{p^*},X^{p^{*\prime}}) \propto C_1$ and $\mathbb{E}_{X^{p^*}}\left[k(X^{p^*},X^{p^*})\right] \leq \sup_{X^{p^*}} k(X^{p^*},X^{p^*}) \propto C_2$.

Further $\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x'\,|\,z')}\frac{p^*(x)}{p(x\,|\,z)}l(y,y')\right] \leq \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\,|\,z')}\right]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\,|\,z)}\right]\sup_{y,y'} l(y,y') = \sup_{y,y'} \propto C_3$. Finally $\mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x\,|\,z)}\right)^2 l(y,y)\right] \leq C_4$ using the same arguments as before (finite variance of the density ratio). Then

$$
\begin{aligned}
\langle B,B\rangle &= \left(\frac{-2}{n}+\frac{1}{n^2}\right)\mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\,|\,z)}\frac{p^*(x')}{p(x'\,|\,z')}l(y,y')\right] \\
&+ \frac{1}{n^2}\mathbb{E}_{X^{p^*}}\left[k(X^{p^*},X^{p^*})\right]\mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x\,|\,z)}\right)^2 l(y,y)\right] \\
&+ \frac{n-1}{n^2}\mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x\,|\,z)}\right)^2 l(y,y)\right] \\
&+ \frac{n-1}{n^2}\mathbb{E}_{X^{p^*}}\left[k(X^{p^*},X^{p^*})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x'\,|\,z')}\frac{p^*(x)}{p(x\,|\,z)}l(y,y')\right] \\
&\leq \left(\frac{-2}{n}+\frac{1}{n^2}\right)C_1C_3 + \frac{1}{n^2}C_2C_4 + \frac{n-1}{n^2}C_1C_4 + \frac{n-1}{n^2}C_2C_3 \propto \mathcal{O}(\frac{1}{n})
\end{aligned}
$$

As $\widehat{C}_{p^*}$ is asymptotically unbiased in L^2 norm, it follows from Chebyshev's inequality that it is a consistent estimator. ■

B.2 Proof for Theorem 2

Proof

Again consider:

$$\mathbb{E}\left[\|C_{p^*} - \widehat{C}_{p^*}\|^2_{\mathrm{HS}}\right]$$

However we take the estimator to be:

$$\widehat{C}_{p^*} = \frac{1}{n}\sum_{i=1}^n \hat{h}_n(x_i,z_i)k(\cdot,x_i)\otimes l(\cdot,y_i) - \left(\frac{1}{n}\sum_{j=1}^n k\left(\cdot,x_j^{p^*}\right)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n \hat{h}_n(x_i,z_i)l(\cdot,y_i)\right).$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg\|\mu_{xy}(\cdot) - \mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot) - \\
&\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^n \hat{h}_n(x_i,z_i)k(\cdot,x_i)\otimes l(\cdot,y_i)}_{\hat{\mu}_{xy}} - \underbrace{\left(\frac{1}{n}\sum_{j=1}^n k\left(\cdot,x_j^{p^*}\right)\right)}_{\hat{\mu}_{X^{p^*}}}\otimes\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \hat{h}_n(x_i,z_i)l(\cdot,y_i)\right)}_{\hat{\mu}_y}\Bigg)\Bigg\|^2\Bigg]
\end{aligned}
$$

We follow the same steps as in Theorem 1.

$$\begin{aligned}&\mathbb{E}\left[\left\|\underbrace{\mu_{xy}(\cdot)-\hat{\mu}_{xy}(\cdot)}_{A}+\underbrace{(\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)-\mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot))}_{B}\right\|^2\right]\\&=\mathbb{E}\left[\langle A,A\rangle+2\langle A,B\rangle+\langle B,B\rangle\right]\end{aligned}$$

First term $\langle A, A\rangle$:

$$\mathbb{E}\left[\langle A,A\rangle\right]=\mathbb{E}\left[\langle\mu_{xy}(\cdot),\mu_{xy}(\cdot)\rangle\right]-2\mathbb{E}\left[\langle\mu_{xy}(\cdot),\hat{\mu}_{xy}(\cdot)\rangle\right]+\mathbb{E}\left[\langle\hat{\mu}_{xy}(\cdot),\hat{\mu}_{xy}(\cdot)\rangle\right]$$

Let $(x', y', z') \sim p$ be an independent copy of $(x, y, z) \sim p$. Then the first term above is:

$$\mathbb{E}\left[\langle\mu_{xy}(\cdot),\mu_{x'y'}(\cdot)\rangle\right]=\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\,|\,z)}\frac{p^*(x')}{p(x'\,|\,z')}k(x,x')l(y,y')\right];$$

the second one is:

$$\begin{aligned}\mathbb{E}\left[\langle\mu_{xy}(\cdot),\hat{\mu}_{xy}(\cdot)\rangle\right]&=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{h}_n(x_i,z_i)\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\,|\,z)}k(x,x_i)l(y,y_i)\right]\right]\\&=\mathbb{E}_{x,y,z,x',y',z'}\left[\hat{h}_n(x,z)\frac{p^*(x')}{p(x'\,|\,z')}k(x,x')l(y,y')\right]\\&=(1+\mathcal{O}(\frac{1}{n^\alpha}))\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\,|\,z)}\frac{p^*(x')}{p(x'\,|\,z')}k(x,x')l(y,y')\right];\end{aligned}$$

and the last term is:

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)k(x_i,x_j)l(y_i,y_j)\right]\\&=\underbrace{\mathbb{E}\left[\frac{1}{n^2}\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)k(x_i,x_j)l(y_i,y_j)\right]}_{(a)}+\underbrace{\mathbb{E}\left[\frac{1}{n^2}\sum_{i}\hat{h}_n(x_i,z_i)^2k(x_i,x_i)l(y_i,y_i)\right]}_{(b)}.\end{aligned}$$

We bound (a):

$$\begin{aligned}(a)&=\frac{(n-1)}{n}\mathbb{E}_{x,y,z,x',y',z'}\left[\hat{h}_n(x,z)\hat{h}_n(x',z')k(x,x')l(y,y')\right]\\&=\frac{(n-1)}{n}\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\,|\,z)}\frac{p^*(x')}{p(x'\,|\,z')}k(x,x')l(y,y')\right]\left(1+2\mathcal{O}(\frac{1}{n^\alpha})+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right)\end{aligned}$$

then (b)

$$
\begin{aligned}
(b) &= \frac{1}{n}\mathbb{E}\left[\left(\hat{h}_n(x,z)\right)^2 k(x,x)l(y,y)\right] \\
&\leq \frac{1}{n}\underbrace{\sup_{x\in X} k(x,x)\sup_{y\in Y} k(y,y)}_{\text{Assumed to be bounded by } C}\underbrace{\mathbb{E}\left[\left(\hat{h}_n(x,z)\right)^2\right]}_{\text{finite variance of density ratio, i.e. bounded by some } D} \\
&= \frac{CD}{n}\left(1 + 2\mathcal{O}(\frac{1}{n^\alpha}) + \mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right)
\end{aligned}
$$

So for $\langle A, A\rangle$ the convergence rate is $\mathcal{O}\left(\frac{1}{n^{\min(1,\alpha)}}\right)$, since that is the slowest decaying term. For the $\langle A, B\rangle$ part we have:

$$
\begin{aligned}
\langle A, B\rangle = &\underbrace{\langle \mu_{xy}(\cdot), \hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(1)} + \underbrace{\langle \hat{\mu}_{xy}(\cdot), \mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(2)} \\
&- \underbrace{\langle \hat{\mu}_{xy}(\cdot), \hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(3)} - \underbrace{\langle \mu_{xy}(\cdot), \mu_{X^{p^*}}(\cdot)\otimes\mu_{y'}(\cdot)\rangle}_{(4)}
\end{aligned}
$$

Each part can then be written as:

$$
\begin{aligned}
(1) &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\hat{h}_n(x_j,z_j)\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}k(x,x_i^{p^*})l(y,y_j)\right]\right] \\
&= \mathbb{E}_{x',y',z'}\left[\hat{h}_n(x',z')\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}k(x,x_i^{p^*})l(y,y_j)\right]\right] \\
&= \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}k(x,x_i^{p^*})l(y,y_j)\right]\right]\left(1+\mathcal{O}\left(\frac{1}{n^\alpha}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
(2) &= \mathbb{E}\left[\frac{1}{n}\sum_{i}^{n}\hat{h}_n(x_i,z_i)\mathbb{E}_{x_q}[k(x_i,X^{p^*})]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}l(y,y_i)\right]\right] \\
&= \mathbb{E}_{x',y',z'}\left[\hat{h}_n(x',z')\mathbb{E}_{x_q}[k(x',X^{p^*})]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}l(y,y')\right]\right] \\
&= \mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\mathbb{E}_{x_q}[k(x',X^{p^*})]\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\mid z)}l(y,y')\right]\right]\left(1+\mathcal{O}\left(\frac{1}{n^\alpha}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
(3) &= \mathbb{E}\left[\frac{1}{n^3}\sum_{i,k,j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)k(x_i,X_k^{p^*})l(y_i,y_j)\right] \\
&= \mathbb{E}\left[\frac{1}{n^3}\sum_{k,i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)k(x_i,X_k^{p^*})l(y_i,y_j)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{n^3}\sum_{k,i=j}\left(\hat{h}_n(x_i,z_i)\right)^2 k(x_i,X_k^{p^*})l(y_i,y_i)\right] \\
&= \left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right)\frac{n-1}{n}\mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\,|\,z')}\right. \\
&\quad \left.\times\mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\,|\,z)}\mathbb{E}_{X^{p^*}}\left[k(x,X^{p^*})\right]l(y,y')\right]\right] \\
&\quad + \left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right)\frac{1}{n}\mathbb{E}_{X^{p^*},x,y,z}\left[\left(\frac{p^*(x)}{p(x\,|\,z)}\right)^2 k(x,X^{p^*})l(y,y)\right].
\end{aligned}
$$

$$
(4) = \mathbb{E}_{x,y,z}\left[\frac{p^*(x)}{p(x\,|\,z)}\mathbb{E}_{x_q}[k(x,X^{p^*})]\mathbb{E}_{x',y',z'}\left[\frac{p^*(x')}{p(x'\,|\,z')}l(y,y')\right]\right]
$$

We can use the same arguments as in $\langle A,A\rangle$, consequently $\langle A,B\rangle = \mathcal{O}\left(\frac{1}{n^{\min(1,\alpha)}}\right)$. The final term $\langle B,B\rangle$ is

$$
\begin{aligned}
\langle B,B\rangle = \underbrace{\langle\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot),\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot)\rangle}_{(1)} - \underbrace{2\langle\hat{\mu}_{X^{p^*}}(\cdot)\otimes\hat{\mu}_y(\cdot),\mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(2)} \\
+ \underbrace{\langle\mu_{X^{p^{*\prime}}}(\cdot)\otimes\mu_{y'}(\cdot),\mu_{X^{p^*}}(\cdot)\otimes\mu_y(\cdot)\rangle}_{(3)},
\end{aligned}
$$

and we have

$$
\begin{aligned}
&(1) = \\
&\mathbb{E}\left[\frac{1}{n^2}\sum_{u,v} k(X_u^{p^*}, X_v^{p^*})\frac{1}{n^2}\sum_{i,j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j)\right] \\
&= \mathbb{E}\Bigg[\frac{1}{n^2}\left(\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*}) + \sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\right) \\
&\quad \frac{1}{n^2}\left(\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j) + \sum_{i=j}\hat{h}_n(x_i,z_i)^2 l(y_i,y_i)\right)\Bigg] \\
&= \underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j)\right]}_{(a)} \\
&\quad + \underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i=j}\hat{h}_n(x_i,z_i)^2 l(y_i,y_i)\right]}_{(b)} \\
&\quad + \underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i=j}\hat{h}_n(x_i,z_i)^2 l(y_i,y_i)\right]}_{(c)} \\
&\quad + \underbrace{\mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j)\right]}_{(d)}
\end{aligned}
$$

Each of these terms is:

$$
\begin{aligned}
(a) &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j)\right] \\
&= \frac{(n-1)^2}{n^2}\mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\frac{p^*(x)}{p(x\mid z)}l(y,y')\right] \\
&\quad\times\left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right) \\
(b) &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i=j}\hat{h}_n(x_i,z_i)^2 l(y_i,y_i)\right] \\
&= \frac{1}{n^2}\mathbb{E}_{X^{p^*}}\left[k(X^{p^*},X^{p^*})\right]\mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x\mid z)}\right)^2 l(y,y)\right] \\
&\quad\times\left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right) \\
(c) &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u\neq v} k(X_u^{p^*}, X_v^{p^*})\sum_{i=j}\hat{h}_n(x_i,z_i)^2 l(y_i,y_i)\right] \\
&= \frac{n-1}{n^2}\mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z}\left[\left(\frac{p^*(x)}{p(x\mid z)}\right)^2 l(y,y)\right] \\
&\quad\times\left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right) \\
(d) &= \mathbb{E}\left[\frac{1}{n^4}\sum_{u=v} k(X_u^{p^*}, X_u^{p^*})\sum_{i\neq j}\hat{h}_n(x_i,z_i)\hat{h}_n(x_j,z_j)l(y_i,y_j)\right] \\
&= \frac{n-1}{n^2}\mathbb{E}_{X^{p^*}}\left[k(X^{p^*},X^{p^*})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x')}{p(x'\mid z')}\frac{p^*(x)}{p(x\mid z)}l(y,y')\right] \\
&\quad\times\left(1+2\mathcal{O}\left(\frac{1}{n^\alpha}\right)+\mathcal{O}\left(\frac{1}{n^{2\alpha}}\right)\right).
\end{aligned}
$$

Returning to the expression for $\langle B, B\rangle$:

$$
\begin{aligned}
(2) &= \mathbb{E}\left[\frac{1}{n}\sum_i \mathbb{E}_{X^{p^*}}[k(X^{p^*},X_i^{p^*})]\frac{1}{n}\sum_j \hat{h}_n(x_j,z_j)\mathbb{E}_{x,y,z}[\frac{p^*(x)}{p(x\mid z)}l(y,y_j)]\right] \\
&= \mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\mid z)}\frac{p^*(x')}{p(x'\mid z')}l(y,y')\right]\left(1+\mathcal{O}\left(\frac{1}{n^\alpha}\right)\right),
\end{aligned}
$$

and

$$
(3) = \mathbb{E}_{X^{p^*},X^{p^{*\prime}}}\left[k(X^{p^*},X^{p^{*\prime}})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\mid z)}\frac{p^*(x')}{p(x'\mid z')}l(y,y')\right].
$$

We use the same arguments as in Theorem 1. We note that the the sums collapse similarly in (1) with some added negligible terms converging faster than $\mathcal{O}(\frac{1}{n})$. What remains is then the slowest converging term

$$\mathbb{E}_{X^{p^*},X^{p^*\prime}}\left[k(X^{p^*},X^{p^*\prime})\right]\mathbb{E}_{x,y,z,x',y',z'}\left[\frac{p^*(x)}{p(x\,|\,z)}\frac{p^*(x')}{p(x'\,|\,z')}l(y,y')\right]\mathcal{O}\left(\frac{1}{n^\alpha}\right)=\mathcal{O}\left(\frac{1}{n^\alpha}\right)$$

compared to $\mathcal{O}(\frac{1}{n})$. Hence $\langle B,B\rangle = \mathcal{O}\left(n^{-\min(1,\alpha)}\right)$. As $\langle A,A\rangle, \langle A,B\rangle, \langle B,B\rangle$ all have the same convergence rate of their bias terms, we conclude that $\mathbb{E}\left[\|C_{p^*}-\widehat{C}_{p^*}\|^2_{\mathrm{HS}}\right] = \mathcal{O}\left(\frac{1}{n^{\min(1,\alpha)}}\right)$. ■

B.3 Proof of Proposition 5

Proof Since $\|\widehat{C}_{p^*}\|^2_{\mathrm{HS}}(\psi)\geq 0$, it suffices by Markov's inequality to show that

$$\lim_{n\to\infty}\mathbb{E}\left[\|\widehat{C}_{p^*}\|^2_{\mathrm{HS}}(\psi)\right]=0.$$

Let let $w_i=\frac{p^*(x_i)}{p(x_i|z_i)}$. Define $\mathbf{M}_k(n):=\{1,\ldots,n\}^k$ as the k-fold Cartesian product of the set $\{1,\ldots,n\}$. We then have

$$\begin{aligned}\|\widehat{C}_{p^*}\|^2_{\mathrm{HS}}(\psi) = &\underbrace{\frac{1}{n^2}\sum_{i,j\in\mathbf{M}_2(n)}k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_iw_j}_{A_n}\\ &+\underbrace{\frac{1}{n^4}\sum_{i,j,k,l\in\mathbf{M}_4(n)}k(x_k,x_l)l(y_{\psi(i)},y_{\psi(j)})w_iw_j}_{B_n}\\ &-2\underbrace{\frac{1}{n^3}\sum_{i,j,k\in\mathbf{M}_3(n)}w_ik(x_i,x_j)w_kl(y_{\psi(i)},y_{\psi(k)})}_{C_n},\end{aligned}\tag{11}$$

which is abbreviated as $A_n+B_n-2C_n$. It suffices to show that $\lim_n\mathbb{E}A_n=\lim_n\mathbb{E}B_n=\lim_n\mathbb{E}C_n=\zeta$. Where

$$\zeta=\mathbb{E}\left[k(X,X^{'})\right]\mathbb{E}\left[l(Y,Y')\right]$$

where X',Z' is an identical independent copy of X,Z respectively. We employ the same strategy as in Rindt et al. (2021) and partition the summing over the indices as

$$A_n=\frac{1}{n^2}\sum_{i,j\in U(2,\psi,n)}k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_iw_j+\frac{1}{n^2}\sum_{i,j\in R(2,\psi,n)}k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_iw_j\tag{12}$$

with

$$U(2,\psi,n):=\{(i,j)\in M_2(n):(i,j,\psi\,(i),\psi\,(j))\text{ are 4 distinct elements}\}$$

and

$$R(2,\psi,n) := M_2(n)\backslash U(2,\psi,n).$$

Here 4 distinct elements simply means that $i, j, \psi(i), \psi(j)$ are all different, i.e. 1,2,3,4. The main observation here is that as $n \to \infty$, almost all terms in sums A_n, B_n, C_n will have distinct terms. We then take

$$\begin{aligned}
\mathbb{E}\left[A_n\right] =& \mathbb{E}\left[\mathbb{E}\left[A_n \mid \psi\right]\right] \\
=& \mathbb{E}\left[\frac{1}{n^2}\sum_{U(2,\psi,n)} \mathbb{E}\left[k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_i w_j \mid \psi\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{n^2}\sum_{R(2,\psi,n)} \mathbb{E}\left[k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_i w_j \mid \psi\right]\right] \\
=& \mathbb{E}\left[\frac{|U(2,\psi,n)|}{n^2}\right]\mathbb{E}\left[k(X,X^{'})\right]\mathbb{E}\left[l(Y,Y')\right] + \mathbb{E}\left[\frac{|R(2,\psi,n)|}{n^2}\right]\mathcal{O}(1) \to \zeta.
\end{aligned} \tag{13}$$

Since the permutation occurs within $p(x \mid z)$, there will always be a dependency between w_i, w_j and $l(y_{\psi(i)}, y_{\psi(j)})$ meaning that distinct indices don't factorize directly unlike the setting in Rindt et al. (2021). However, by calculating the expectation we have that

$$\begin{aligned}
&\mathbb{E}\left[k(x_i,x_j)l(y_{\psi(i)},y_{\psi(j)})w_i w_j \mid \psi\right] \\
&= \mathbb{E}\left[k(X,X')l(Y,Y')WW'\right] \\
&= \int k(x,x')l(y,y')\frac{p^*(x)}{p(x \mid z)}\frac{p^*(x')}{p(x' \mid z')}p(x,y,z)p(x',y',z')dxdydzdx'dy'dz' \\
&= \int k(x,x')l(y,y')p^*(x)p^*(x')\underbrace{\left(\int p(y \mid x,z)p(z)dz\right)\left(\int p(y' \mid x',z')p(z')dz'\right)}_{y \text{ and } x \text{ are rendered independent by having distinct indices}}dxdydx'dy' \\
&= \int k(x,x')l(y,y')p^*(x)p^*(x')p(y)p(y')dxdx'dydy' \\
&= \mathbb{E}\left[k(X,X^{'})\right]\mathbb{E}\left[l(Y,Y')\right] = \zeta.
\end{aligned} \tag{14}$$

Now we can repeat the argument in Rindt et al. (2021) with

$$\begin{aligned}
\frac{\mathbb{E}|U(2,\psi,n)|}{n^2} &= \frac{n(n-1)}{n^2}\cdot\mathbb{P}\left((i,j,\psi(i),\psi(j)) \text{ are 4 distinct elements }\right) \\
&= \frac{n(n-1)}{n^2}\frac{\binom{n-2}{2}\binom{n-4}{2}\cdots\binom{n-2d+2}{2}}{\binom{n}{2}\binom{n}{2}\cdots\binom{n}{2}} \\
&\longrightarrow 1.
\end{aligned} \tag{15}$$

Hence $\lim_{n\to\infty}\mathbb{E}(A_n) = \mathbb{E}[k(X,X')]\,\mathbb{E}[l(Y,Y')] = \zeta$. We can repeat the argument for B_n and C_n and hence the sum of ζ's collapse into 0. Throughout the proof, we have assumed that $\mathbb{E}[k(X,X')l(Y,Y')WW'] < C$, where C is some constant. ■

Appendix C. Proof of Proposition 2

We first establish some "RKHS calculus" before we proceed with the calculations.

Mean Some rules following Riesz representation theorem and RKHS spaces.

1. $\langle \mu_x, f\rangle_{\mathcal{F}} = \mathbf{E}_x[\langle \phi(x), f\rangle_{\mathcal{F}}] = \mathbf{E}_x[f(x)]$
2. $\langle \mu_y, g\rangle_{\mathcal{G}} = \mathbf{E}_y[\langle \psi(y), g\rangle_{\mathcal{G}}] = \mathbf{E}_y[g(x)]$
3. $\|\mu_x\|^2_{\mathcal{F}} = \mathbf{E}_{x,x'}[\langle \phi(x), \phi(x')\rangle_{\mathcal{F}}] = \mathbf{E}_{x,x'}[k(x,x')] = \frac{1}{n^2}\sum_{i,j} k(x_i, x_j)$

Tensor operator $\otimes$

We may employ any $f \in \mathcal{F}$ and $g \in \mathcal{G}$ to define a tensor product operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ as follows: $(f \otimes g)h := f\langle g, h\rangle_{\mathcal{G}}$ for all $h \in \mathcal{G}$

Lemma 1. *For any $f_1, f_2 \in \mathcal{F}$ and $g_1, g_2 \in \mathcal{G}$ the following equation holds:* $\langle f_1 \otimes g_1, f_2 \otimes g_2\rangle_{\mathrm{HS}} = \langle f_1, f_2\rangle_F \langle g_1, g_2\rangle_G$

Using this lemma, one can simply show the norm of $f \otimes g$ equals $\|f \otimes g\|^2_{\mathrm{HS}} = \|f\|^2_{\mathcal{F}}\|g\|^2_{\mathcal{G}}$

Let $\hat{\mu}_{P_x} = \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \cdot)$ and $\hat{\mu}_{P_y} = \frac{1}{n}\sum_{i=1}^n l(\mathbf{y}_i, \cdot)$. We are now ready to proceed with the derivation.

Proof For

$$\widehat{C}_{p^*} = \frac{1}{n}\sum_{i=1}^n \tilde{w}_i k(\cdot, x_i) \otimes l(\cdot, y_i) - \left(\frac{1}{m_x}\sum_{j=1}^{m_x} k\left(\cdot, x_j^{p^*}\right)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \tilde{w}_i l\left(\cdot, y_i\right)\right).$$

we have the following:

$$\|\widehat{C}_{p^*}\|^2 = A + B - 2C$$

where

$$A = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \tilde{w}_i\tilde{w}_j \langle k(\cdot, x_i) \otimes l(\cdot, y_i), k(\cdot, x_j) \otimes l(\cdot, y_j)\rangle = \tilde{w}^\top (K \circ L)\,\tilde{w} = tr\left(D_{\tilde{w}} K D_{\tilde{w}} L\right),$$

$$B = \frac{1}{m_x^2 n^2}\left\|\sum_{i=1}^n k\left(\cdot, x_i^{p^*}\right)\right\|^2 \left\|\sum_{i=1}^n \tilde{w}_i l\left(\cdot, y_i\right)\right\|^2 = \frac{1}{m_x^2 n^2}(\mathbf{K}^{p^*})_{++}(\mathbf{L} \circ \tilde{W})_{++}$$

$$
\begin{aligned}
C &= \frac{1}{n^2 m_x} \left\langle \sum_{i=1}^{n} \tilde{w}_i k(\cdot, x_i) \otimes l(\cdot, y_i), \left(\sum_{j=1}^{m_x} k(\cdot, x_j^{p^*}) \right) \otimes \left(\sum_{j=1}^{n} \tilde{w}_j l(\cdot, y_j) \right) \right\rangle \\
&= \frac{1}{n^2 m_x} \sum_{i=1}^{n} \tilde{w}_i \left(\sum_{j=1}^{m_x} k(x_i, x_j^{p^*}) \right) \left(\sum_{r=1}^{n} \tilde{w}_r l(y_i, y_r) \right) \\
&= \tilde{w}^\top \left(K^q 1_{m_x} \circ L\tilde{w} \right) \\
&= \operatorname{tr}\left(D_{\tilde{w}} K^q 1_{m_x} \tilde{w}^\top L \right),
\end{aligned}
$$

It should be noted that we can choose the number of samples m_x for $x_i^{p^*} \sim p^*$ to use. For practical purposes we set $m_x = n$. ■

Appendix D. Optimal c_{p^*} Choice

D.1 Derivation of univariate c_{p^*}

Suppose that we wish to choose an 'optimal' c_{p^*} value for rescaling the distribution. One criterion for optimality is to maximize the effective sample size

$$
\mathrm{ESS} := \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}
$$

where $w_i = p^*\left(x; \phi^*\right) / p_{X|Z}(x; \phi)$ is the weight used to resample observations. This is essentially equivalent to minimizing the variance of the individual weights:

$$
\arg\min_{\phi \cdot} \operatorname{Var} w\left(X_i, Z_i; \phi^*\right).
$$

Note that

$$
\begin{aligned}
\mathbb{E} w\left(X_i, Z_i; \phi^*\right) &= \int \frac{p^*\left(x, z, \phi^*\right)}{p_{X|Z}(x, z, \phi)} p_{X|Z}(x, z, \phi) dx dz \\
&= \int p^*\left(x, z, \phi^*\right) dx dz \\
&= 1
\end{aligned}
$$

so this is equivalent to minimizing the squared expectation of $w()$. If we assume that everything is Gaussian and that X, Z have standard normal marginal distributions with correlation ρ, this becomes equivalent to minimizing

$$
\frac{1}{\tau^2} \mathbb{E} \left[\frac{\phi\left(\frac{X}{\tau}\right)}{\phi\left(\frac{X - \rho Z}{\sqrt{1-\rho^2}}\right)} \right]^2
$$

with respect to τ. We can rewrite this expression as:

$$\begin{aligned} f(\tau) &= \frac{1}{2\pi\sqrt{1-\rho^2\tau^2}} \iint_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\begin{pmatrix} x \\ z \end{pmatrix}^T \begin{pmatrix} \frac{2}{\tau^2} - \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1-2\rho^2}{1-\rho^2} \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix}\right\} dxdz \\ &= \frac{1}{\sqrt{1-\rho^2\tau^2}} \begin{vmatrix} \frac{2}{\tau^2} - \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1-2\rho^2}{1-\rho^2} \end{vmatrix}^{-1/2} \\ &= \frac{1}{\tau^2} \begin{vmatrix} \frac{2(1-\rho^2)}{\tau^2} - 1 & \rho \\ \rho & 1-2\rho^2 \end{vmatrix}^{-1/2} \end{aligned}$$

Note that minimizing f is the same as maximizing $1/f^2$, so we need to maximize

$$\begin{aligned} 1/f(\tau)^2 &= \tau^4 \begin{vmatrix} \frac{2(1-\rho^2)}{\tau^2} - 1 & \rho \\ \rho & 1-2\rho^2 \end{vmatrix} \\ &= \tau^4 \left(\left(2\frac{1-\rho^2}{\tau^2} - 1\right)(1-2\rho^2) - \rho^2\right)\right) \\ &= \tau^2 2\left(1-\rho^2\right)\left(1-2\rho^2\right) - \left(1-\rho^2\right)\tau^4. \end{aligned}$$

This is maximized at $\tau^2 = 1-2\rho^2$, or $\tau = \sqrt{1-2\rho^2}$.

D.2 Derivation of multivariate c_{p^*}

Now suppose that $(X, Z) \sim N_{p+p^*}(0, \Sigma)$ where we take $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}$ and assume that $\Sigma_{xx} = I_p$ and $\Sigma_{zz} = I_{p^*}$ (again, this can be achieved by rescaling). By the same reasoning as above, we want to minimize the squared expectation of the weights, which amounts to minimizing

$$\begin{aligned} f(\tau) =& \frac{1}{|T|} \iint_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\begin{pmatrix} x \\ z \end{pmatrix}^T \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix}\right\} dx\, dz \\ &\propto \frac{1}{|T|} \begin{vmatrix} A & B \\ C & D \end{vmatrix}^{-1/2} \end{aligned}$$

where

$$\begin{aligned} A &= 2T^{-1} - (I_p - \Sigma_{xz}\Sigma_{zx})^{-1}, \\ B &= (I_p - \Sigma_{xz}\Sigma_{zx})^{-1}\Sigma_{xz}, \\ C &= \Sigma_{zx}(I_p - \Sigma_{xz}\Sigma_{zx})^{-1}, \\ D &= I_{p^*} - \Sigma_{zx}(I_p - \Sigma_{xz}\Sigma_{zx})^{-1}\Sigma_{xz}. \end{aligned}$$

or maximizing

$$|T|^2 \begin{vmatrix} 2T^{-1} - (I_p - \Sigma_{xz}\Sigma_{zx})^{-1} & (I_p - \Sigma_{xz}\Sigma_{zx})^{-1}\Sigma_{xz} \\ \sum_{zx}(I_p - \Sigma_{xz}\Sigma_{zx})^{-1} & I_{p^*} - \Sigma_{zx}(I_p - \Sigma_{xz}\Sigma_{zx})^{-1}\Sigma_{xz} \end{vmatrix}$$

Note that in this case we will choose a whole matrix T, rather than just a scaling constant, but we could simplify to assume that $T = c_{p^*} I_p$ for some scalar c_{p^*}.

We take
$A := 2T^{-1} - (I_p - \Sigma_{xz}\Sigma_{zx})^{-1}$, $B := (I_p - \Sigma_{xz}\Sigma_{zx})^{-1} \Sigma_{xz}$, $D := I_{p^*} - \Sigma_{zx} (I_p - \Sigma_{xz}\Sigma_{zx})^{-1} \Sigma_{xz}$.
The block matrix is then

$$M := \left| \begin{array}{cc} A & B \\ B^\top & D \end{array} \right|$$

Assuming D is invertible, we can use the Schur complement

$$\det(M) = \det(D) \det\left(A - BD^{-1}B^\top\right)$$

Then

$$\det(A - BD^{-1}B^\top) = \det(2T^{-1} - (I_p - \Sigma_{xz}\Sigma_{zx})^{-1} - BD^{-1}B^\top).$$

which can easily be optimized with gradient descent with respect to $T = c_{p^*} I_p$. For estimated covariance matrices $\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{xx} & \hat{\Sigma}_{xz} \\ \hat{\Sigma}_{zx} & \hat{\Sigma}_{zz} \end{bmatrix}$. It suffices to plug them in directly in the estimator.

Appendix E. bd-CME test derivation

We observe $\{(x_i, y_i, z_i)\}_{i=1}^n \sim p$, some probability density on the joint space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Define

$$p(y|do(x)) \propto \tilde{p}(y|do(x)) = \int p(y|x,z)p(z)dz.$$

Note that $p(y|do(x)) \neq p(y|x)$ in general since X and Z need not be independent. We would like to test

$$H_0 : \ p(y|do(x)) \text{ does not depend on } x$$

versus the general alternative. Let k, l, m be kernels on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. We first fit the conditional mean embedding

$$\mu_{p(\cdot|x,z)} \quad = \quad \int l(\cdot, y)p(y|x,z)dy,$$

by learning a regression function $k(\cdot, x) \otimes m(\cdot, z) \mapsto l(\cdot, y)$. This is given by $\hat{\mu}_{p(\cdot|x,z)} = \sum_{i=1}^n a(x,z)_i l(\cdot, y_i)$, where

$$a(x,z) = \left(K_{\mathbf{xx}} \circ M_{\mathbf{zz}} + \epsilon_n I\right)^{-1} \left(K_{\mathbf{x}x} \circ M_{\mathbf{z}z}\right),$$

$[K_{\mathbf{xx}}]_{ij} = k(x_i, x_j)$ is the $n \times n$ Gram matrix and $K_{\mathbf{x}x} \in \mathbb{R}^n$, with $[K_{\mathbf{x}x}]_i = k(x_i, x)$, and similarly for kernel m. Now, to obtain an estimate of

$$\mu_{\tilde{p}(\cdot|do(x))} \quad = \quad \int l(\cdot, y) \int p(y|x,z)p(z)dzdy,$$

we simply average over the empirical $\hat{p}(z)$ to obtain:

$$\hat{\mu}_{\tilde{p}(\cdot|do(x))} = w(x)^\top L_{\mathbf{y}\cdot} = \sum_{i=1}^{n} w(x)_i l(\cdot, y_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a(x, z_j)_i l(\cdot, y_i),$$

i.e.

$$w(x) = \frac{1}{n}\left(K_{\mathbf{xx}} \circ M_{\mathbf{zz}} + \epsilon_n I\right)^{-1}\left(K_{\mathbf{x}x} \circ M_{\mathbf{zz}}\mathbf{1}\right).$$

Denote

$$W_{\mathbf{x}} = [w(x_1)\cdots w(x_n)]^\top = \frac{1}{n}\left(K_{\mathbf{xx}} \circ \mathbf{1}\mathbf{1}^\top M_{\mathbf{zz}}\right)\left(K_{\mathbf{xx}} \circ M_{\mathbf{zz}} + \epsilon_n I\right)^{-1}.$$

We also denote $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w(x_i)$. Then $\hat{\mu}_{\tilde{p}} = \bar{w}^\top L_{\mathbf{y}\cdot}$ estimates the embedding of $\tilde{p}(y) = \int p(y|x,z)p(x)p(z)dxdz$. Now, under the null, $\tilde{p} = \tilde{p}(\cdot|do(x)), \forall x$. Thus, we define the statistic as the sum of the RKHS distances of the corresponding embeddings

$$\begin{aligned}
S &= \sum_{i=1}^{n} \left\|\hat{\mu}_{\tilde{p}(\cdot|do(x)=x_i)} - \hat{\mu}_{\tilde{p}}\right\|^2_{\mathcal{H}_l} \\
&= \sum_{i=1}^{n} \left\|\left(w(x_i) - \bar{w}\right)^\top L_{\mathbf{y}\cdot}\right\|^2_{\mathcal{H}_l} \\
&= \sum_{i=1}^{n} \mathrm{Tr}\left[\left(w(x_i) - \bar{w}\right)^\top L_{\mathbf{yy}}\left(w(x_i) - \bar{w}\right)\right] \\
&= \mathrm{Tr}\left[L_{\mathbf{yy}}\sum_{i=1}^{n}\left(w(x_i) - \bar{w}\right)\left(w(x_i) - \bar{w}\right)^\top\right] \\
&= \mathrm{Tr}\left[L_{\mathbf{yy}} W_{\mathbf{x}}^\top H W_{\mathbf{x}}\right].
\end{aligned}$$

The test statistic S uses a mean embedding that bears many similarities to Singh et al. (2023). In this context, we consider an RBF kernel on Y, which takes the outcome to the RKHS.

Appendix F. Simulation algorithms

F.1 Binary Treatment

Fix $\beta_{XY} > 0$, and $\tau^2 > 0$. H_0 case:

$$\begin{aligned}
Z_i &\sim \mathcal{N}(0,1), \\
X_i|Z_i &\sim \text{Bernoulli}\left(\frac{1}{1+e^{-Z_i}}\right), \\
Y_i|Z_i &\sim \mathcal{N}\left(\beta_{XY} Z_i, \tau^2\right).
\end{aligned}$$

H_1 case:

$$\begin{aligned} Z_i &\sim \mathcal{N}(0,1), \\ X_i|Z_i &\sim \text{Bernoulli}\left(\frac{1}{1+e^{-Z_i}}\right), \\ Y_i|Z_i &\sim \mathcal{N}\left(\beta_{XY}\left(2X_i-1\right)|Z_i|, \tau^2\right). \end{aligned}$$

Note that in the alternative case, X_i directly modulates the sign of the mean of Y_i. However, $2X_i - 1$ will be strongly positively correlated with the sign of Z_i implying that there is only a slight change in the dependence structure of (X_i, Y_i, Z_i). In addition, all the marginals are the same.

F.2 Continuous treatment

We simulate data from the do-null and the alternative using rejection sampling (Evans and Didelez, 2024), described in Algorithm 5. To construct a data set where there is marginal dependence (Figure 2a) but the do-null holds, we replace the normal marginal distributions for X, Y, Z in Algorithm 5 with exponential distributions, i.e. $Y \sim \text{Exp}(X\beta_{XY})$. To generate data where there is conditional dependence but the do-null holds, we simply use the last set of parameters in Appendix G with normal marginal distributions.

Algorithm 5: Generating continuous data for H_0 and H_1

Input: Number of samples n, dependencies β_{XY}, β_{XZ}, β_{YZ}, variance parameters $\theta, \phi \in \mathbb{R}^+$, dimensions d_x, d_y, d_z

Initialize data container $\mathcal{D} = \{\}$

Set $\boldsymbol{\beta}_{XZ} = [\underbrace{\beta_{XZ}}_{1:3}, \underbrace{0}_{4:d_Z}]$

while # *of samples* $< n$ **do**

- Set $p_X = \mathcal{N}(\mathbf{0}, \theta \cdot \phi \cdot \mathbf{I}_{d_x})$ Sample $\{x_i\}_{i=1}^N \sim p_X$
- Sample $\{y_i, z_i\}_{i=1}^N \sim \mathcal{N}(\mathbf{0}, \Sigma_{d_y+d_z})$
- Transform $\{y_i'\}_{i=1}^N = \text{CDF}_{\mathcal{N}(0,1)}(\{y_i\}_{i=1}^N)$
- Define $p_{Y|X} = \mathcal{N}(X\beta_Y, 1)$
- Set $\{y_i\}_{i=1}^N = \text{ICDF}_{p_{Y|X}}(\{y_i'\}_{i=1}^N)$
- Set $\mu_{X|Z} = Z \cdot \boldsymbol{\beta}_{XZ}$
- Define $p_{X|Z} = \mathcal{N}(\mu_{X|Z}, \phi)$
- Calculate $\omega_i = \frac{p_{X|Z}(x_i)}{p_X(x_i)}$
- Run rejection sampling using ω_i and obtain $\mathcal{D}' = \{x_i, y_i, z_i\}_{i=1}^{N'} \sim p^*$
- Append data $\mathcal{D} = \mathcal{D} \cup \mathcal{D}'$

end

Return: $\mathcal{D}$

F.2.1 Parameter explanation

There are several parameters used in the data generation algorithm primarily used to control for the difficulty of the problem and the ground truth hypothesis.

1. β_{XY}: Controls the dependency between X and Y. A $\beta_{XY} > 0$ implies H_1 ground truth and $\beta_{XY} = 0$ implies H_0 ground truth
2. β_{XZ}: Controls the dependency between X and Z. A high β_{XZ} implies a stronger dependency on Z for X, implying a harder problem.
3. β_{YZ}: Controls the dependency between Y and Z. Fixed at a high value to ensure Y is being confounded by Z.
4. θ, ϕ: Controls the variance of p_X and $p_{X|Z}$. $\theta > \phi$ ensures a higher *Effective Sample Size* for true weights.
5. d_x, d_y, d_z: Dimensionality of X, Y, Z. Higher dimensions imply a harder problem.
6. $\Sigma_{d_y+d_z}$: The covariance matrix that links Z and Y. It should be noted that this covariance matrix is a function of X.

F.3 Mixed treatment

Algorithm 6: Generating mixed data for H_0 and H_1

Input: Number of samples n, dependencies β_{XY}, β_{XZ}, β_{YZ}, variance parameters $\theta, \phi \in \mathbb{R}^+$, dimensions d_x, d_y, d_z

Initialize data container $\mathcal{D} = \{\}$

Set $\boldsymbol{\beta}_{XZ} = [\underbrace{\beta_{XZ}}_{1:3}, \underbrace{0}_{4:d_Z}]$

while # *of samples* $< n$ **do**

- Set $p_X = \mathcal{N}(\mathbf{0}, \theta \cdot \phi \cdot \mathbf{I}_{d_x})$
- Set $p_X^{\text{bin}} = \text{Bin}(p = 0.5)$
- Sample $\{x_i^{\text{cont}}\}_{i=1}^N \sim p_X$
- Sample $\{x_i^{\text{bin}}\}_{i=1}^N \sim \text{Bin}(p = 0.5)$
- Concatenate $X = X_{\text{cont}} \cup X_{\text{bin}}$
- Sample $\{y_i, z_i\}_{i=1}^N \sim \mathcal{N}(\mathbf{0}, \Sigma_{d_y+d_z})$
- Transform $\{y_i'\}_{i=1}^N = \text{CDF}_{\mathcal{N}(0,1)}(\{y_i\}_{i=1}^N)$
- Define $p_{Y|X} = \mathcal{N}(X\beta_Y, 1)$
- Set $\{y_i\}_{i=1}^N = \text{ICDF}_{p_{Y|X}}(\{y_i'\}_{i=1}^N)$
- Set $\mu_{X|Z} = Z \cdot \boldsymbol{\beta}_{XZ}$
- Set $\nu_{X|Z} = \frac{1}{1+e^{-\mu_{X|Z}}}$
- Define $p_{X|Z} = \mathcal{N}(\mu_{X|Z}, \phi)$
- Define $p_{X|Z}^{\text{bin}} = \text{Bin}(p = \nu_{X|Z})$
- Calculate $\omega_i = \frac{p_{X|Z}(x_i^{\text{cont}})}{p_X(x_i^{\text{cont}})} \cdot \frac{p_{X|Z}^{\text{bin}}(x_i^{\text{bin}})}{p_X^{\text{bin}}(x_i^{\text{bin}})}$
- Run rejection sampling using ω_i and obtain $\mathcal{D}' = \{x_i, y_i, z_i\}_{i=1}^{N'} \sim p^*$
- Append data $\mathcal{D} = \mathcal{D} \cup \mathcal{D}'$

end

Return: $\mathcal{D}$

Appendix G. Parameters for data generation

We provide parameters used in the data generation procedure for each type of treatment. Exact details can be found in the code base. *Binary treatment*

1. Dependency β_{XY}: $[0.0, 0.02, 0.04, 0.06, 0.08, 0.1]$
2. Variance τ: 1.0

Continuous treatment

1. β_{XY}: $[0.0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.008, 0.012, 0.016, 0.02]$
2. β_{XZ}: $d_Z = 1 : 0.75, d_Z = 3, 15, 50 : 0.25$
3. β_{YZ}: $[0.5, 0.0]$
4. θ, ϕ: $d_Z = 1 : (2, 2)$, $d_Z = 3 : (4, 2)$, $d_Z = 15 : (8, 2)$, $d_Z = 50 : (16, 2)$
5. $\Sigma_{d_y+d_z}$: See code base for details

Mixed treatment

1. β_{XY}: $[0.0, 0.002, 0.004, 0.006, 0.008, 0.01, 0.015, 0.02, 0.025, 0.03, 0.04, 0.05, 0.1]$
2. β_{XZ}: 0.05
3. β_{YZ}: $[0.5, 0.0]$
4. θ, ϕ: $d_Z = 2 : (2, 2)$, $d_Z = 15 : (16, 2)$, $d_Z = 50 : (16, 2)$
5. $\Sigma_{d_y+d_z}$: See code base for details

$X \not\perp Y$ *data*

1. β_{XY}: $[0.0]$
2. β_{XZ}: 1.0
3. β_{YZ}: $[0.5, 0.0]$
4. θ, ϕ: $d_Z = 1 : (0.1, 1.5)$
5. $\Sigma_{d_y+d_z}$: See code base for details

$X \not\perp Y \mid Z$ *data*

1. β_{XY}: $[0.0]$
2. β_{XZ}: 0.0.
3. β_{YZ}: $[-0.5, 4.0]$
4. θ, ϕ: $d_Z = 1 : (1.0, 2.0)$
5. $\Sigma_{d_y+d_z}$: See code base for details

References

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650, 04 2014. ISSN 0034-6527. doi: 10.1093/restud/rdt044. URL https://doi.org/10.1093/restud/rdt044.

Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957, 2010. doi: 10.1214/10-AOS799. URL https://doi.org/10.1214/10-AOS799.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1), Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL http://dx.doi.org/10.1038/s41467-020-17478-w.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Rhian M Daniel, SN Cousens, BL De Stavola, Michael G Kenward, and JAC Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618, 2013.

Laurent Charlin Dawen Liang and David M. Blei. Causal inference for recommendation. In *RecSys '20: Fourteenth ACM Conference on Recommender Systems*, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141. Citeseer, 2014.

Tim Ensor, Stephanie L. Cooper, Lisa Davidson, Ann E. Fitzmaurice, and Wendy Jane Graham. The impact of economic recession on maternal and infant mortality: lessons from history. *BMC Public Health*, 10:727 – 727, 2010.

Robin J Evans and Vanessa Didelez. Parameterizing and simulating from causal models (with discussion), 2024. to appear, arXiv:2109.03694.

Jake Fawkes, Robert Hu, Robin J. Evans, and Dino Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects even under one sided overlap, 2022. URL https://arxiv.org/abs/2212.04922.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, page 63–77, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 354029242X. doi: 10.1007/11564089_7. URL https://doi.org/10.1007/11564089_7.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012. URL http://jmlr.org/papers/v13/gutmann12a.html.

Vasant Honavar and Sanghack Lee. A kernel conditional independence test for relational data. *Conference on Uncertainty in Artificial Intelligence*, 01 2017.

Alexander P Keil, Jessie P Buckley, Katie M O'Brien, Kelly K Ferguson, Shanshan Zhao, and Alexandra J White. A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environmental health perspectives*, 128(4):047004, 2020.

Scott Kostyshak. Non-parametric testing of U-shaped relationships. *Econometrics: Econometric & Statistical Methods - General eJournal*, 2017.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning, 2023.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

Taku Moriyama and Masashi Kuwano. Causal inference for contemporaneous effects and its application to tourism product sales data. *Journal of Marketing Analytics*, 08 2021. doi: 10.1057/s41270-021-00130-x.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/rezende15.html.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4905–4916. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/33d3b157ddc0896addfb22fa2a519097-Paper.pdf.

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. 2023.

David Rindt, Dino Sejdinovic, and David Steinsaltz. A kernel- and optimal transport- based test of independence between covariates and right-censored lifetimes. *The International Journal of Biostatistics*, page 20200022, 2020. doi: doi:10.1515/ijb-2020-0022. URL https://doi.org/10.1515/ijb-2020-0022.

David Rindt, Dino Sejdinovic, and David Steinsaltz. Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC. *Stat*, 10(1):e364, 2021. doi: https://doi.org/10.1002/sta4.364. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.364. e364 sta4.364.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ISSN 00063444. URL http://www.jstor.org/stable/2335942.

Kenneth J. Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150, 2005. doi: 10.2105/AJPH.2004.059204. URL https://doi.org/10.2105/AJPH.2004.059204. PMID: 16030331.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Bernhard Scholköpf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

R Singh, L Xu, and A Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, page asad042, 07 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad042. URL https://doi.org/10.1093/biomet/asad042.

Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 823–830, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273600. URL https://doi.org/10.1145/1273496.1273600.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019. doi: doi:10.1515/jci-2018-0017. URL https://doi.org/10.1515/jci-2018-0017.

Daiki Watanabe, Tsukasa Yoshida, Yuya Watanabe, Yosuke Yamada, and Misaka Kimura. A U-shaped relationship between the prevalence of frailty and body mass index in community-dwelling japanese older adults: The kyoto–kameoka study. *Journal of Clinical Medicine*, 9, 2020.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/d1f255a373a3cef72e03aa9d980c7eca-Paper.pdf.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, Jan 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9721-7. URL http://dx.doi.org/10.1007/s11222-016-9721-7.