A Kernel Two-Sample Test with Selection Bias

Alexis Bellot[1,2] **Mihaela van der Schaar**[1,2,3]

[1]University of Cambridge, Cambridge, UK.
[2]Alan Turing Institute, London, UK
[3]University of California, Los Angeles, Los Angeles, USA

Abstract

Hypothesis testing can help decision-making by quantifying distributional differences between two populations from observational data. However, these tests may inherit biases embedded in the data collection mechanism (some instances often being systematically more likely included in our sample) and consistently reproduce biased decisions. We propose a two-sample test that adjusts for selection bias by accounting for differences in marginal distributions of confounding variables. Our test statistic is a weighted distance between samples embedded in a reproducing kernel Hilbert space, whose balancing weights provably correct for bias. We establish the asymptotic distributions under null and alternative hypotheses, and prove the consistency of empirical approximations to the underlying population quantity. We conclude with performance evaluations on artificial data and experiments on treatment effect studies from economics.

1 INTRODUCTION

The two-sample problem considers testing whether two independent samples are likely drawn from the same distribution. Such tests have a long history in statistical inference but they are also increasingly used in decision making scenarios. For example, two-sample tests have been used to determine gender differences in academic achievements (Hyde, 2005), gender differences in criminal justice outcomes (Grabe *et al.*, 2006), gender differences in health issues (Verbrugge, 1985), and also frequently used in medicine to determine subgroups of patients that respond differently to medication and establish treatment policies (Biesecker, 2013).

In any data driven study, a *first* step is the collection of a series of observations about an underlying phenomenon of interest before making an informed decision, for example assisted by a hypothesis test, on this data. In most realistic scenarios, we do not have control on the data collection process (e.g. participants volunteering for a study involving a new treatment may differ systematically from the wider population), but we do implicitly condition on the fact that participants entered into the study $(S = 1)$.

This implicit conditioning may *bias* the conclusions of tests because two samples may differ systematically prior to running any experiment and a hypothetical difference in distribution be completely unrelated to the effect of interest.

The problem of selection bias and its influence on inference has attracted much recent interest in the fairness literature (Pearl, 2012; John *et al.*, 2020; Kilbertus *et al.*, 2020; Doroudi *et al.*, 2017), one aspect of which involves mitigating indirect discrimination e.g., section 3.1, point (2) in (Zliobaite, 2015), in which algorithms make biased decisions due to the correlation of the non-discriminatory items with the discriminatory ones. Selection bias is also relevant in the causal inference literature (Pearl, 2000). (Bareinboim & Pearl, 2012) gave graphical conditions under which the causal effect may be recovered from data with selection bias. A similar scenario is considered under the rubric of treatment effect estimation, in which algorithms estimate individualized, average and conditional treatment effects in data biased by confounders that simultaneously influence treatment assignment and outcomes (Wager & Athey, 2018; Johansson *et al.*, 2016; Zhang *et al.*, 2020). In epidemiology (Robins *et al.*, 2000) and econometrics (Heckman, 1979), versions of this problem are also widely studied. Similarly, hypothesis tests for the significance of a measured association, and data-driven algorithms in general, must account for sources of discrimination, confounding, and selection bias more generally in the data.

In fairness and causal inference however, while many methods exist attempting to predict associations adjusting for selection bias, much less is known on the *significance* of effects in the presence of selection bias. We cannot for instance say whether outcome distributions in two groups

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:205–214.

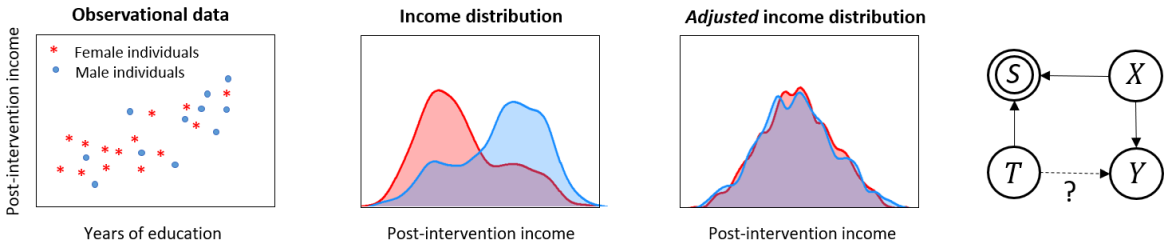

Figure 1: **The influence of selection bias on outcome distributions**. The left panel plots a sample from the observed data, the middle panel shows the observed post-intervention income density for males and females, and the right panel shows the income distribution obtained by adjusting for selection bias, in this case education levels. In this case, a conventional two-sample test rejects the hypothesis of equal post-intervention income in male and female populations due to the intervention, while our proposed test fails to reject.

significantly differ or not even if model predictions differ. The literature on hypothesis testing is invested in such problems but so far hypothesis testing in the presence of selection bias has been minimally considered. To illustrate the large impact of selection bias on two-sample testing outcomes and the need for approaches that can adjust for these spurious effects we consider an example described below and illustrated in Figure 1.

1.1 EXAMPLE

Suppose a city government wants to understand the role of gender on the effectiveness of a past employment program to better allocate their resources in the future. Its analyst constructed datasets of volunteering ($S = 1$) men and women ($T = 1$ and $T = 0$) to be compared, and included a number of relevant employment figures such as post-intervention earnings, type of job, satisfaction, etc. (Y). In this hypothetical example, highly educated men were more likely to volunteer than women due to historical gender bias in education opportunities (X). Such preferential selection creates a *spurious* association between T and Y, opening a path of unblocked correlations through X, as shown in the causal diagram of Figure 1. It is called spurious because it is not part of what we seek to estimate – **the significance of the causal effect of T on Y**.

A test that ignores this bias tends to determine men and women to have different employment program outcomes whereas in reality, once we account for differences in education (i.e. we block the spurious open path), the program is found to perform equally in distribution across men and women. In this example, higher program benefits are due to higher starting education standards, not because people of different gender benefit differently. A decision based on a plain two sample tests overrates the impact of an individual's sex – in this case correlated with education because we implicitly condition on $S = 1$. Please find a description of the data generating mechanism in the Appendix.

1.2 CONTRIBUTIONS

We develop a non-parametric test for differences in distribution of two samples biased by preferential selection driven by other observed quantities.

Our proposal is a generalization of two sample tests based on maximum mean discrepancies between probability distributions (Gretton *et al.*, 2012; Chwialkowski *et al.*, 2015; Jitkrittum *et al.*, 2017; Zaremba *et al.*, 2013; Bellot & van der Schaar, 2019) that incorporate importance sampling techniques to adjust for distributional shift in covariates. The technical challenge is that adjustments made for differences in the marginal confounding distributions between two samples are data-dependent, and therefore invalidate existing asymptotic guarantees of tests based on the maximum mean discrepancy.

Our contributions are three-fold.

1. We propose a two-sample test statistic that, under certain conditions, provably adjusts for selection bias.
2. We derive novel asymptotic distributions for the proposed test.
3. In the finite-sample case, we propose weight approximations for our test statistic, that we show to be consistent with its population-level quantity.

2 BACKGROUND

From the context of hypothesis testing, to understand the role of selection bias it is useful to bring in knowledge of the causal mechanisms in data and augment a causal graph with a variable S that represents the recruitment of individuals into the study. The assignment of individuals into two groups $T \in \{0, 1\}$ is then correlated with confounding variables $X \in \mathcal{X}$ through the fact that we condition on individuals to be included in the study (see Figure 1). We call these confounding variables because they introduce spurious differences in the relationship between outcome

variables and the selection mechanism once we condition on $S = 1$. To formalise hypothesis testing with biased data, we adopt the potential outcomes framework of (Rubin, 2005). We assume to have observed independent samples from an outcome variable $Y = Y^1 \cdot T + Y^0 \cdot (1 - T)$, the response variable Y is split into counterfactual variables, Y^0 and Y^1, had $T = 0$ and $T = 1$ occurred respectively, i.e. under a model where selection bias does not influence treatment assignment.

The hypothesis testing problem is formulated as evaluating the evidence for a difference in distribution P_{Y^1} and P_{Y^0} in two groups of observations,

$$\mathcal{H}_0 : P_{Y^1} = P_{Y^0} \quad \text{versus} \quad \mathcal{H}_1 : P_{Y^1} \neq P_{Y^0}, \tag{1}$$

but, unlike conventional two-sample problems, we have access to distributions P_{Y^1} and P_{Y^0} only via an (unknown) sampling policy $T \in \{0, 1\}$ that introduces bias due to the implicit conditioning on $S = 1$, rather than directly through independent samples from P_{Y^1} and P_{Y^0}. S and T create distributional shift, the assumption is that the available data is independently sampled from *distorted* distributions conditional on T. The counterfactual distributions P_{Y^0} and P_{Y^1} we are interested in differentiating are not directly observed. Instead, through available samples we have access to $P_{Y|T=0}$ and $P_{Y|T=1}$, different from P_{Y^0} and P_{Y^1} because $(Y^1, Y^0) \not\perp\!\!\!\perp T|S = 1$. The same attributes X that correlate with the probability of group assignment T may also be associated with the potential responses Y^0 and Y^1.

2.1 PRELIMINARIES ON HYPOTHESIS TESTING

The problem of hypothesis testing is to define a test statistic (a function of observational data) to distinguish between two hypotheses on the distribution of observed samples. Short of perfectly distinguishing between any two hypotheses we may pose due to the limited number of samples available to characterize distributions, tests are constructed such that a certain hypothesis is rejected whenever a test statistic exceeds a certain threshold away from 0 (Lehmann & Romano, 2006). The goal of hypothesis testing is to derive a threshold such that false positives are upper bounded by a design parameter α and false negatives are as low as possible.

Our test statistic is characterized by distances in mean embeddings of distributions in a reproducing kernel Hilbert space $\mathcal{H}_k$. The advantage of mapping distributions P_{Y^0} and P_{Y^1} to functions in $\mathcal{H}_k$ is that we may now say that P_{Y^0} and P_{Y^1} are close if the RKHS distance $||\mu_{P_{Y^0}} - \mu_{P_{Y^1}}||_{\mathcal{H}_k}$ is small, where $\mu_P := \int_{\mathcal{X}} k(x, \cdot) dP(x)$ is the embedding of the probability measure P to $\mathcal{H}_k$. This distance is known as the Maximum Mean Discrepancy (MMD) (Gretton *et al.*, 2012) and is particularly appealing because for certain choices of the kernel function k, the mean embedding can be shown to be injective (Sriperumbudur *et al.*, 2011). All properties of the distribution are conserved with this map and one may distinguish between distributions by computing the MMD between them.

$$\text{MMD}(P_{Y^0}, P_{Y^1}) = 0 \quad \text{if and only if} \quad P_{Y^0} = P_{Y^1}. \tag{2}$$

We focus our attention on the Gaussian kernel $k(x, y) = \exp(-||x - y||^2/\sigma^2)$ with bandwidth parameter σ, that enjoys this property. The squared MMD is given by (Gretton *et al.*, 2012),

$$\begin{aligned}\text{MMD}^2 := &\mathop{\mathbb{E}}_{y,y^\star \sim P_{Y^1}} k(y, y^\star) + \mathop{\mathbb{E}}_{y,y^\star \sim P_{Y^0}} k(y, y^\star) \\ &- 2 \mathop{\mathbb{E}}_{y \sim P_{Y^1}, y^\star \sim P_{Y^0}} k(y, y^\star),\end{aligned} \tag{3}$$

and empirical estimates may be computed in practice.

3 AN IMPORTANCE WEIGHTED STATISTIC

With access only to samples from biased populations $P_{Y|T=1}$ and $P_{Y|T=0}$ estimating the above distance with respect to counterfactual distributions P_{Y^0} and P_{Y^1} empirically is not possible. To ensure identifiability of the hypothesis testing problem however, we may assume that (Y^0, Y^1) and the data generating process satisfy ignorability: $Y^0, Y^1 \perp\!\!\!\perp T|X, S = 1$, a common assumption in the treatment effect estimation literature. It means that within any stratum of X, individuals who would have one set of potential outcomes $Y(0) = y_0$ and $Y(1) = y_1$, are just as likely to be in the control or treatment group as other individuals (with different potential outcomes) that share characteristics X. If in addition we assume that $0 < Pr(T|X) < 1$, then with knowledge of the sample selection mechanisms $e(x) := Pr(T = 1|X = x)$ we may recover the expectations of interest with importance sampling,

$$\begin{aligned}\mathbb{E}\left(\frac{Y}{e(X)} \,\middle|\, T = 1\right) &= \mathbb{E}\left(\frac{T \cdot Y^1}{e(X)}\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\frac{T \cdot Y^1}{e(X)} \,\middle|\, X\right)\right) \\ &= \mathbb{E}\left(Y^1\right).\end{aligned} \tag{4}$$

This encourages us to define a weighted estimator of the MMD - called the WMMD - such that the weights emphasize distances in areas of the support where the distributions of confounding variables agree. Define w such that $Pr(T = 1|X = x) \cdot w(x) = Pr(T = 0|X = x)$ and consider,

$$\begin{aligned}\text{WMMD}^2 := &\mathop{\mathbb{E}}_{(x,y),(x^\star,y^\star) \sim P_{XY|T=1}} w(x)w(x^\star)k(y, y^\star) \\ &+ \mathop{\mathbb{E}}_{y,y^\star \sim P_{Y|T=0}} k(y, y^\star) \qquad (5) \\ &- 2 \mathop{\mathbb{E}}_{\substack{(x,y) \sim P_{XY|T=1}, \\ y^\star \sim P_{Y|T=0}}} w(x)k(y, y^\star) \qquad (6)\end{aligned}$$

where the superscript $\star$ denotes an independent copy where appropriate. We show next that this metric consistently distinguishes between null and alternative hypotheses at the population level.

Proposition 1 *For k a characteristic kernel and known positive weights, WMMD $= 0$ if and only if $P_{Y_1} = P_{Y^0}$.*

Proof. All proofs are given in the Appendix.

A kernel k is characteristic if the mean embedding μ_P is injective.

In practice, we have access to an empirical estimate of the WMMD, defined as follows,

$$\begin{aligned}\widehat{\text{WMMD}}^2 := &\sum_{i\neq j: t_i=t_j=1} w_i w_j k(y_i, y_j) \\ &+ \sum_{i\neq j: t_i=t_j=0} k(y_i, y_j) \\ &- 2 \sum_{i,j: t_i=1, t_j=0} w_i k(y_i, y_j),\end{aligned}$$

where the (y_i, t_i, x_i) are realizations of the random variables (Y, T, X) and where we have written $w_i = w(x_i)$ for brevity (we will use these two notations interchangeably). Deviations from 0 (the theoretical value under the null) are expected due to finite sample variation. Tests are then constructed such that the null hypothesis is rejected whenever $\widehat{\text{WMMD}}^2$ exceeds a certain threshold. In the next section we will show how to consistently define such a threshold to ensure a low margin of error.

3.1 HYPOTHESIS TESTING WITH WMMD

As we have mentioned, from the statistical testing point of view, the coincidence axiom of the WMMD is key, as it ensures consistency against any alternative hypothesis $\mathcal{H}_1$. Then, given a significance level α for the two-sample test, a test can be constructed such that $\mathcal{H}_0$ is rejected when $\widehat{\text{WMMD}}^2 > r$.

The expected behaviour of $\widehat{\text{WMMD}}^2$ under the null which we might use to define r however differs from conventional bounds used for U-statistics. The reason is that in practice weights are data-dependent and have their own asymptotic behaviour which needs to be accounted for. In this case, under mild conditions that ensure well defined limits for these weights, also the asymptotic distributions are well defined. This result is given in Theorem 1 below.

Theorem 1 (Asymptotic distribution of WMMD). *Assume that k has finite second moments and that the weight matrix $W \in \mathbb{R}^{n\times n}$ ($W_{ij} = w_i w_j$) be approximately diagonalizable (made precise in the Appendix). Then, the following statements hold,*

1. *Under $\mathcal{H}_0$, the asymptotic distribution of $\widehat{WMMD}^2$ is given by a mixture of independent χ^2 random variables.*
2. *Under $\mathcal{H}_1$,*

$$n^{1/2}\left(\widehat{WMMD}^2 - WMMD^2\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right).$$

For conciseness, we have omitted here the exact terms of the scaled χ^2 distribution and asymptotic variance which are intricate but can be found in the Appendix. We have used $\xrightarrow{d}$ to denote convergence in distribution.

3.2 APPROXIMATING THE WEIGHTS IN PRACTICE

While we have shown that our test statistic is consistent against all alternatives, in practice simulating from the asymptotic null distribution can be challenging. The distribution under the null requires knowledge of the sample selection mechanism, that is the design densities of the assignment variable T in the two populations, which is not available.

A straightforward solution is to estimate each function $Pr(T = 1|X = x)$ and $Pr(T = 0|X = x)$ separately, for example with a classification algorithm, although this has been shown to result in unstable estimates of the ratio $Pr(T = 1|X = x)/Pr(T = 0|X = x)$ when the denominator is small (Sugiyama *et al.*, 2007) and adds an additional computational burden to the test procedure. An alternative approach is to use a plug-in estimate for the ratio directly. The approach we take is to estimate weights $\hat{w}(x)$ such that $Pr(T = 1|X = x) \approx \hat{w}(x)Pr(T = 0|X = x)$ by matching feature representation of both domains in a high-dimensional feature space (Gretton *et al.*, 2009).

We estimate weights $\hat{w}$ such as to minimize the distance between mean embeddings in a RKHS $\mathcal{H}_K$ with kernel K that is defined by a feature map $\phi : \mathcal{X} \to \mathcal{H}_K$ of the confounding variable distributions in the two populations,

$$\hat{w} := \underset{0<w<B}{\text{argmin}} \left|\left|\mathbb{E}_{P_{X|T=0}} w(x)\phi(x) - E_{P_{X|T=1}}\phi(x)\right|\right|_{\mathcal{H}_K}. \tag{7}$$

This problem is convex. For injective mappings, minimizing (7) converges to $Pr(T = 1|X = x)/Pr(T = 0|X = x)$ and $\hat{w}$ can be found with a quadratic program for which many efficient solvers have been developed (Diamond & Boyd, 2016). In our implementation (see more details in the Appendix) we use the Gaussian kernel with bandwidth parameter set to the median Euclidian distance between values of the confounding variables. Theorem 2 below guarantees that the density ratio estimation using (7) in the computation of $\widehat{\text{WMMD}}$ and of the asymptotic null distribution still yields a consistent test.

Theorem 2 (Consistency of $\widehat{\text{WMMD}}$). *Let $\hat{w}(x)$ be the empirical density ratio estimates of $w(x)$ - the underlying pop-*

ulation value - derived by matching the kernel mean embeddings of the observed distributions of confounding variables $P_{X|T=1}$ and $P_{X|T=0}$. Suppose the test threshold is set to the upper α quantile of the distribution of the WMMD under $\mathcal{H}_0$. Then, asymptotically, the false positive rate with estimated weights is α and its power converges to 1.

The proof, given in the Appendix, is based on the consistency of kernel mean matching to approximate the likelihood ratio in the asymptotic regime. While importance weighting using the likelihood ratio results in $\widehat{\text{WMMD}}$ being an asymptotically unbiased estimator of the MMD, the estimator may not concentrate well because the weights may be large or inaccurate due to the finite samples available in practice. We now provide a concentration bound for $\widehat{\text{WMMD}}$ for the case where weights are upper-bounded by some maximum value.

Theorem 3 (Large deviation bound of $\widehat{\text{WMMD}}$). *Let $\{y_i, t_i, x_i\}_{i=1}^{n+m}$ be i.i.d observations drawn from the joint distribution of random variables (Y, T, X), n of them with $t_i = 1$ and m with $t_i = 0$. Assume the feature representation $\phi(x) \in H_\phi$ to have maximum value R, $w(x) \leq B$ for all $x \in \mathcal{X}$, and that there exists an $\epsilon > 0$ such that,*

$$\left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}(x_i)\phi(x_i) - \frac{1}{m} \sum_{i=1:t_i=0}^{m} \phi(x_i) \right\|_{H_\phi} \leq \epsilon.$$

Then, with probability at least $1 - \delta$, the absolute difference in estimation of weighted estimator $\widehat{WMMD}$ in comparison to the MMD, $|\widehat{WMMD}^2 - MMD^2|$ is bounded above by,

$$2R(B+1)\left(\epsilon + \left(1 + \sqrt{2\log\frac{2}{\delta}}\right) R\sqrt{\frac{B^2}{n} + \frac{1}{m}}\right) + R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}},$$

where $m_2 := \lfloor m/2 \rfloor$.

Qualitatively, B measures the maximum allowed discrepancy between $Pr(T = 1|X = x)$ and $Pr(T = 0|X = x)$ (and is a user defined parameter in practice, we set it to 10 as a default in our experiments). A low value of B ensures robustness of the learned representations by limiting the influence of individual observations, thus reducing the variance of the resulting estimator and improving its concentration around the true estimate. However, with strong bias the discrepancy between $Pr(T = 1|X = x)$ and $Pr(T = 0|X = x)$ is large and limiting B will result in higher ϵ which increases the bound. In turn, as expected, concentration improves with sample size. Asymptotically in m and n with high probability, the concentration of the representation depends only on matching confounding distributions in feature space ϕ. This shows that unbiased two-sample testing is not possible unless enough *comparable* examples in the two populations exist.

3.3 CONNECTIONS WITH TESTING IN REGRESSION MODELS

There is a close connection between testing for distributional differences in two outcome samples independent of confounding and the predictive power of those factors on the outcome. In fact, adjustment is needed precisely because confounding variables are both predictive of the outcome and predictive of the sample selection mechanism. In one approach, the source of variation due to sample selection bias on the outcome y can be modelled explicitly, for example by considering a regression model with random effects. Consider the following random effect regression model (Schall, 1991) for the outcome y,

$$Y_i = \mu + Z_i u_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{8}$$

where $Z_i \in \{0, 1\}$ represents the assignment of example i into one of the two samples and $u_i \sim \mathcal{N}(0, \sigma_u^2)$. Under the null assumption, testing for variation in Y that is irrelevant of the sample selection mechanism (which is our goal) is then equivalent to testing the variance component $\sigma_u^2 = 0$ (Goeman *et al.*, 2004; Lin, 1997). A score test statistic for this problem is given by $S = \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}\tilde{Y}_i\tilde{Y}_j + \sum_{i=1}^{n}\tilde{Y}_i^2$ where $\tilde{Y}_i := \frac{(Y_i - \mu)}{\sigma}$, see e.g Section 4 in (Goeman *et al.*, 2004). The statistic S therefore has a high value whenever the terms of the matrix $K = (k_{ij})$ and the matrix $\tilde{Y}\tilde{Y}^T$ with (i, j)-th element $(\tilde{Y}_i\tilde{Y}_j)$ are correlated. Now consider the case $n = m$ and write $y_{i,1} = y_i$ if $t_i = 1$, and analogously for $y_{j,0}$, $i, j = 1, \ldots, n$. Let k_{ij} be a column vector with entries $[k(y_{i,1}, y_{j,1}), k(y_{i,0}, y_{j,0}), k(y_{i,1}, y_{j,0}), k(y_{i,0}, y_{j,1})]$ and let w_{ij} have entries $[w(x_i)w(x_j), 1, -w(x_i), -w(x_j)]$. Then we may write,

$$\widehat{\text{WMMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} w_{ij}^T k_{ij},$$

which can be interpreted as a non-linear alternative to the first term of S where the inner product $\langle a, b\rangle = a^T b$ is replaced by the inner product in feature space $k(a, b)$.

4 RELATED WORK

Many empirical studies, especially those investigating treatments and effects from finite samples require a notion of statistical significance to assess treatment outcomes.

Existing proposals test for significance of estimated parameters in a regression model and are mostly concerned with average effects or average effects within defined subgroups (Crump *et al.*, 2008; Ding *et al.*, 2014) and not with differences in the outcome distribution as a whole as considered in this paper. To our knowledge no hypothesis test exist for the two-sample problem in the presence of selection bias.

Existing tests, in some cases, may be adjusted to accommodate for selection bias. One possible approach is using ANCOVA (Analysis of Covariance) methods which proceed by regressing the outcome variable on confounding variables before comparing the variation of the corresponding residuals between the two populations to the variation of the residuals within each one of the two populations, for example with an F-test (Tabachnick *et al.*, 2019). Another approach is use a non-parametric alternative, for example using power series as basis functions as proposed in (Crump *et al.*, 2008). However, in these cases, the hypothesis being tested tends to be restricted to average effects.

One extension to (full distribution) two-sample testing that may be considered for this problem is to first, partition the combined population into homogeneous subgroups (such that the feature distribution of confounding variables approximately agree in each subgroup, for example using the propensity score) and second, compute two sample tests statistics in each subgroup before averaging their results. Such tests would take the form of block tests or B-tests (Zaremba *et al.*, 2013), proposed initially as more efficient alternatives to conventional tests. In our experiments, we implement non-parametric versions of each one of these, see the Appendix for a more detailed description.

Outside the hypothesis testing literature, weighted statistics are frequent, often referred to as importance sampling techniques and inverse probability weighting methods (Sugiyama *et al.*, 2007). Using importance weights with the MMD specifically has been used in generative models to sample from modified distributions (Diesendruck *et al.*, 2018) and for unsupervised domain adaptation (Yan *et al.*, 2017, 2019).

5 EXPERIMENTS

In this section we compare two-sample tests on both artificial benchmark data and real-world data. The focus of our results will be on the evaluation of **power**: the rate at which we correctly reject $\mathcal{H}_0$ when it is false; and **type I error**: the rate at which we incorrectly reject $\mathcal{H}_0$ when it is true. $\alpha = 0.05$ throughout.

Baseline Tests. The proposed test is denoted **WMMD**. Comparisons are made with three tests. The **ANCOVA** F-test based on regression residuals from a random forest model. The block-based approach where partitions are made based on the propensity score and two-sample tests in each partition conducted with the MMD (Zaremba *et al.*, 2013) (**Block-MMD**). The Block-MMD can be seen as an alternative adjusting for selection bias in subsets of the data separately, rather than continuously as with our approach and which we expect to have uncontrolled type I error in heterogeneous data samples. And finally, the unweighted (conventional) **MMD** test (Gretton *et al.*, 2012) that serves to measure the benefit of adjustments for selection bias as well as any loss in performance by using the WMMD in data that is not biased.

For kernel-based tests, since their null distributions are given by an infinite sum of weighted chi-squared variables (no closed-form quantiles), in each trial we use 400 random permutations to approximate the null. Details of implementations are given in the Appendix.

5.1 SYNTHETIC EXAMPLES

The primary objective of our synthetic simulations will be to analyse the influence of the sampling selection mechanism on performance. Here it will be particularly interesting to understand our test's behaviour on samples that appear different (in distribution) but only because of an underlying mismatch in confounding variables that simultaneously influence the distributions of interest. In this case we would expect conventional two sample tests to reject the null hypothesis resulting in uncontrolled type I error ($> \alpha$). And similarly for the case of observed distributions that seem to match (in distribution) due to spurious correlations that we show results in low power of traditional tests.

Experiment design. We consider the following data distributions for two samples of data ($T = 0$ and $T = 1$) that exhibit a spurious dependence between their respective outcome distributions $Y|T = 0$ and $Y|T = 1$ such as might occur due to selection bias,

$$X|T = 0 \sim \mathcal{N}(0, I), \qquad X|T = 1 \sim \mathcal{N}(\mu, \sigma^2 I),$$
$$Y|T = i \sim g_i(X) + \mathcal{N}(0, I), \quad i = 0, 1.$$

With this data generating mechanism, units in our two samples ($T = 0$ and $T = 1$) have differing confounder distributions $X|T$, a systematic difference which creates a spurious connection between T and Y.

Recall that the hypothesis testing problem is to evaluate, with data sampled from the model above, the evidence for a difference in distribution P_{Y^1} and P_{Y^0},

$$\mathcal{H}_0 : P_{Y^1} = P_{Y^0} \quad \text{versus} \quad \mathcal{H}_1 : P_{Y^1} \neq P_{Y^0}. \tag{9}$$

μ and σ^2 determine selection bias, i.e. the extent of the dependence between X and T which biases the dependence between T and Y. The distributions we are interested in discriminating are P_{Y^0} and P_{Y^1} (which reduces to $g_0 = g_1$ under the null, and $g_0 \neq g_1$ under the alternative), which implicitly remove selection bias by breaking the dependency between X and T.

5.1.1 Performance with increasing bias

In a first experiment we investigate the influence of increasing selection bias with two problems:

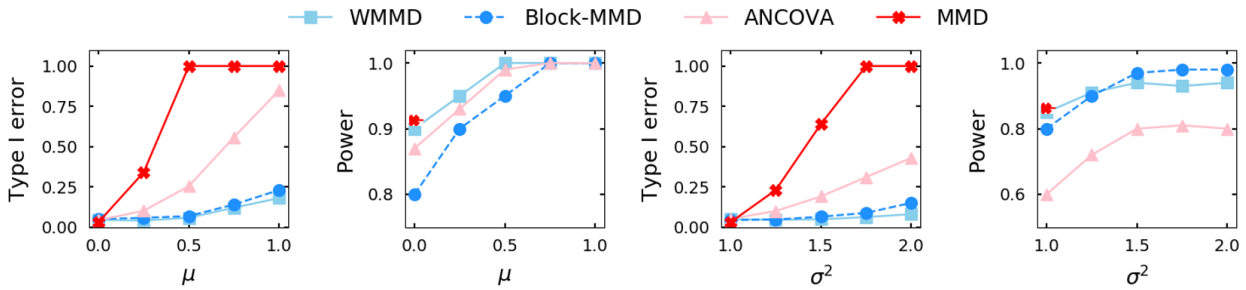

Figure 2: Type I error (lower better) and Power (higher better) of all tests on synthetic experiments. The proposed test is **WMMD**. The WMMD has simultaneously best control of type I errors and highest power.

1. Difference in means μ (with $\sigma^2 = 1$) of confounding variables across the two samples. Results in the two left-most panels of Figure 2.

2. Difference in variances σ^2 (with $\mu = 0$) of confounding variables across the two samples. Results in the two right-most panels of Figure 2.

In each case the dimensionality of X and Y are set to 20, the number of samples in each population to $n = 400$. Under $\mathcal{H}_0$, $g_0(x) = g_1(x) = x + x^2$, and under $\mathcal{H}_1$, $g_0(x) = x$ and $g_1(x) = [\sin(x_1), x_2, ..., x_{20}]$. This set-up is designed to be a challenging problem with moderately high-dimensionality, non-linear dependencies and for the alternative hypothesis differences only in the first dimension of X.

Results. Across experiments (Figure 2) WMMD is the only test that successfully adjusts for selection bias, with controlled type I error even in relatively high bias settings (for instance for $\mu = 1$, only 60% of their densities overlap) while other alternatives underperform.

As anticipated, conventional two-sample test such as the MMD fail with the presence of confounders, we omit plotting the MMD for the power results (beyond $\mu = 0$ and $\sigma^2 = 1$) due to its poor type I error control. We notice also that the Type I error of the block-MMD deteriorates substantially for the variance experiment, potentially because a coarse partition may introduce artificial differences between samples that lead the test to reject the null more often than desired. The panels describing power show good performance for all methods. It is also expected that power increases with confounder distributional shift, as it results in more divergent outcome distributions (and thus easier to distinguish). However, unless type I error is controlled, those results lose their significance. Among methods that control type I error (WMMD and Block-MMD for low bias settings i.e. first half of each panel approximately), WMMD has higher or competitive power.

We make an important comparison also in the two power experiments in the absence of selection bias (the point where the MMD in red is computed). The MMD and WMMD have comparable performance, which suggests that the WMMD is *almost as efficient* as the MMD in datasets tailored to the latter (when no bias exists), while also having good performance in the presence of bias. This is important because in most cases it is not known which variables confound the association between group membership and outcome. What this result means is that we are not worse-off using the WMMD even when there is no selection bias. In this sense the WMMD generalizes the MMD.

5.1.2 Relating to our theoretical results

Even though performing competitively, we observe the WMMD to loosen control of type I error as the strength of bias increases. In the following experiments we consider data generated under $\mathcal{H}_0$ as described in the first paragraph of section 5.1.1. and investigate the estimated WMMD statistic in comparison with optimal behaviour (defined as "True MMD", that is the MMD computed from data with no unobserved confounding on distributions Y^0, Y^1 not accessible in practice).

Results. With increasing confounding, we see in the left-most panel of Figure 3 that the WMMD departs from its optimal value. The reason is that matching distributions of confounders gets harder with increasing confounding. Notice for instance the increasing value of ϵ in the opposite vertical axis, that quantifies the difference between matched distributions introduced in Theorem 3. The middle panel shows however that this discrepancy rapidly vanishes with increasing sample size. Here, we have fixed $\mu = 1$ and increased the sample size to see the estimation error converging to zero.

The takeaway is that a larger number of samples can be expected to be required to successfully control for type I errors to the desired threshold, while the number of samples depends on the strength of the confounding bias among the two samples.

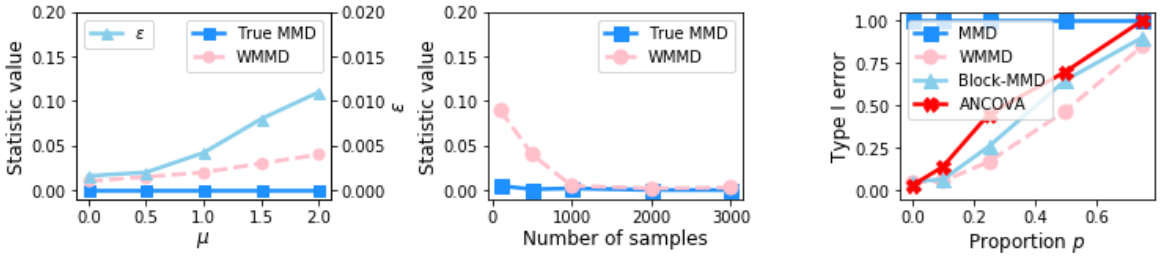

Figure 3: The two leftmost panels show the approximation error of the WMMD with increasing bias and increasing sample size - see details in section 5.1.2. The rightmost panel show type I errors in the presence of unobserved confounders - see details in section 5.1.3.

5.1.3 What if confounding is unobserved?

We have assumed until now that the selection bias is completely driven by factors available to the researcher. In most real applications this will not be the case. We simulate such a scenario by including unobserved confounders in the sample selection mechanism under the null with the same specifications considered above. To do so, we hide or remove from the observed data a proportion p of variables X.

Results. The results, as a function of p, are shown in the rightmost panel of Figure 3. Unobserved confounders introduce variation in the outcome distribution that cannot be adjusted for since it is unobserved, which translates in uncontrolled type I errors for all methods. One may not expect to consistent hypothesis testing in this scenario but we note that this criticism extends to all methods with an assumption of ignorability, and in particular including most treatment effect estimation algorithms.

Remark. Variables X, treated as confounders in our case, may play other roles in general graphical models, for example as mediators or colliders (in both cases with an arrow from T into X). For the purposes of two-sample testing of treatment effects however we may rule out both of these cases because of temporal precedence, i.e. we cannot have an arrow going from T into X because group (treatment) assignment is done *after* observation of X. In others, if T represents a pre-existing characteristic of individuals (such as gender in the in the example of the introduction) we must validate the causal graph to ensure correct conclusions.

5.2 EMPLOYMENT PROGRAM EVALUATION

The problem is to determine the effectiveness of an employment program implemented in the mid-1970s in the U.S. to individuals who had faced economic and social hardship (LaLonde, 1986). The outcome of interest is earnings two years after the end of the employment program. Our null hypothesis is no difference in earnings with the program, with respect to earnings without the program. Posterior earning in treated and control populations are not directly comparable because the populations differ systematically in their education level, prior earnings, age, ethnicity and marital status: all plausible confounders. The data contains 614 individuals, 185 of whom were included in the employment program.

p	**0.05**	**0.10**	**0.15**	**0.20**
MMD	0.95	1	1	1
Block-MMD	0.051	0.055	0.070	0.083
ANCOVA	0.045	0.040	0.056	0.096
WMMD	0.051	0.043	0.052	0.060

Table 1: Type I error at level $\alpha = 0.05$ as a function of artificially introduced bias p.

Experiment design. With real data, the ground truth relationship between two populations is unknown. To compare the performance of our test, however, we can simulate a distribution under the null $\mathcal{H}_0$ by shuffling all variables into two populations, and subsequently introducing bias by selectively removing observations based on a set of confounding covariates. To remove observations, we build a linear regression model to predict earnings based on confounding variables and remove those observations with *high* predicted earnings in one group and those with *low* predicted earning in the other group. After adjusting for this bias the two populations should be equal in distribution and performance comparisons are then made in terms of type I error. A similar approach is used for conventional two sample testing, see for example the experiments in (Lopes *et al.*, 2011).

Results. Type I error as a proportion p of observations removed (that is increasing bias) is given in Table 1. On the original data, all tests returned significant difference in earnings with and without the employment program. This is an important result in its own right as it demonstrates an effect independent of selection bias.

6 CONCLUSIONS

We have proposed a test statistic for the two-sample problem that expands the toolkit of statisticians to make inference on treatment effects with selection-biased data. Bias in the sample selection mechanism creates distributional shift which leads to bias in the treatment effect if unaccounted for. Making inference on the significance of treatment effects in this context is challenging and under-explored. To our knowledge, our test is the first to consider two-sample testing in biased groups of data.

Our proposal is a generalization of the MMD to adjust for this bias. We have demonstrated our test to be consistent in the presence of selection bias, derived its asymptotic distribution and derived large deviation bounds of approximations in practice. In empirical comparisons, we have shown our test to be more powerful than existing alternatives while controlling approximately for type I error.

The weighting strategy and proof techniques presented in this paper are not specific to the two sample problem and may be applied to kernel-based tests for other problems, such as independence testing (Gretton *et al.*, 2007), conditional independence testing (Zhang *et al.*, 2012) and three variable interaction testing (Sejdinovic *et al.*, 2013). Similarly, one may extend the proposed approach to test and adjust for selection bias in other structured spaces where kernels are known to be characteristic such as other compact metric spaces (Bellot & van der Schaar, 2019).

Acknowledgements

We thank the anonymous reviewers for valuable feedback. This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, the ONR and the NSF grants number 1462245 and number 1533983.

REFERENCES

Bareinboim, Elias, & Pearl, Judea. 2012. Controlling selection bias in causal inference. *Pages 100–108 of: Artificial Intelligence and Statistics*.

Bellot, Alexis, & van der Schaar, Mihaela. 2019. *Kernel Hypothesis Testing with Set-valued Data*.

Biesecker, Leslie G. 2013. Hypothesis-generating research and predictive medicine. *Genome research*, **23**(7), 1051–1053.

Chwialkowski, Kacper P, Ramdas, Aaditya, Sejdinovic, Dino, & Gretton, Arthur. 2015. Fast two-sample testing with analytic representations of probability measures. *Pages 1981–1989 of: Advances in Neural Information Processing Systems*.

Crump, Richard K, Hotz, V Joseph, Imbens, Guido W, & Mitnik, Oscar A. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, **90**(3), 389–405.

Diamond, Steven, & Boyd, Stephen. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, **17**(1), 2909–2913.

Diesendruck, Maurice, Cole, Guy W, & Williamson, Sinead. 2018. Directing Generative Networks with Weighted Maximum Mean Discrepancy.

Ding, Peng, Feller, Avi, & Miratrix, Luke. 2014. Randomization inference for treatment effect variation. *arXiv preprint arXiv:1412.5000*.

Doroudi, Shayan, Thomas, Philip S, & Brunskill, Emma. 2017. Importance Sampling for Fair Policy Selection. *Grantee Submission*.

Goeman, Jelle J, Van De Geer, Sara A, De Kort, Floor, & Van Houwelingen, Hans C. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–99.

Grabe, Maria Elizabeth, Trager, KD, Lear, Melissa, & Rauch, Jennifer. 2006. Gender in crime news: A case study test of the chivalry hypothesis. *Mass Communication & Society*, **9**(2), 137–163.

Gretton, Arthur, Borgwardt, Karsten, Rasch, Malte, Schölkopf, Bernhard, & Smola, Alex J. 2007. A kernel method for the two-sample-problem. *Pages 513–520 of: Advances in neural information processing systems*.

Gretton, Arthur, Smola, Alex, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten, & Schölkopf, Bernhard. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning*.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, & Smola, Alexander. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, **13**(Mar), 723–773.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Hyde, Janet Shibley. 2005. The gender similarities hypothesis. *American psychologist*, **60**(6), 581.

Jitkrittum, Wittawat, Xu, Wenkai, Szabó, Zoltán, Fukumizu, Kenji, & Gretton, Arthur. 2017. A linear-time kernel goodness-of-fit test. *Pages 262–271 of: Advances in Neural Information Processing Systems*.

Johansson, Fredrik, Shalit, Uri, & Sontag, David. 2016. Learning representations for counterfactual inference. *Pages 3020–3029 of: International conference on machine learning*.

John, Philips George, Vijaykeerthy, Deepak, & Saha, Diptikalyan. 2020. Verifying individual fairness in machine learning models. *Pages 749–758 of: Conference on Uncertainty in Artificial Intelligence*. PMLR.

Kilbertus, Niki, Ball, Philip J, Kusner, Matt J, Weller, Adrian, & Silva, Ricardo. 2020. The sensitivity of counterfactual fairness to unmeasured confounding. *Pages 616–626 of: Uncertainty in Artificial Intelligence*. PMLR.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Lehmann, Erich L, & Romano, Joseph P. 2006. *Testing statistical hypotheses*. Springer Science & Business Media.

Lin, Xihong. 1997. Variance component testing in generalised linear models with random effects. *Biometrika*, **84**(2), 309–326.

Lopes, Miles, Jacob, Laurent, & Wainwright, Martin J. 2011. A more powerful two-sample test in high dimensions using random projection. *Pages 1206–1214 of: Advances in Neural Information Processing Systems*.

Pearl, Judea. 2000. *Causality: models, reasoning and inference*. Vol. 29. Springer.

Pearl, Judea. 2012. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*.

Robins, James M, Hernan, Miguel Angel, & Brumback, Babette. 2000. *Marginal structural models and causal inference in epidemiology*.

Rubin, Donald B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**(469), 322–331.

Schall, Robert. 1991. Estimation in generalized linear models with random effects. *Biometrika*, **78**(4), 719–727.

Sejdinovic, Dino, Gretton, Arthur, & Bergsma, Wicher. 2013. A kernel test for three-variable interactions. *arXiv preprint arXiv:1306.2281*.

Sriperumbudur, Bharath K, Fukumizu, Kenji, & Lanckriet, Gert RG. 2011. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, **12**(Jul), 2389–2410.

Sugiyama, Masashi, *et al.* 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 985–1005.

Tabachnick, Barbara G, Fidell, Linda S, & Ullman, Jodie B. 2019. *Using multivariate statistics*. Vol. 7. Pearson Boston, MA.

Verbrugge, Lois M. 1985. Gender and health: an update on hypotheses and evidence. *Journal of health and social behavior*, 156–182.

Wager, Stefan, & Athey, Susan. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, **113**(523), 1228–1242.

Yan, Hongliang, Ding, Yukang, Li, Peihua, Wang, Qilong, Xu, Yong, & Zuo, Wangmeng. 2017. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *Pages 2272–2281 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yan, Hongliang, Li, Zhetao, Wang, Qilong, Li, Peihua, Xu, Yong, & Zuo, Wangmeng. 2019. Weighted and Class-Specific Maximum Mean Discrepancy for Unsupervised Domain Adaptation. *IEEE Transactions on Multimedia*, **22**(9), 2420–2433.

Zaremba, Wojciech, Gretton, Arthur, & Blaschko, Matthew. 2013. B-test: A non-parametric, low variance kernel two-sample test. *Pages 755–763 of: Advances in neural information processing systems*.

Zhang, Kun, Peters, Jonas, Janzing, Dominik, & Schölkopf, Bernhard. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zhang, Yao, Bellot, Alexis, & van der Schaar, Mihaela. 2020. Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754*.

Zliobaite, Indre. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*.