A Kernel Two-Sample Test for Unbiased Decisions

Alexis Bellot[1,2], **Mihaela van der Schaar**[1,2,3]
[1]University of Cambridge, [2]The Alan Turing Institute, [3]University of California Los Angeles
[abellot,mschaar]@turing.ac.uk

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, including in scenarios with implications for individuals and society [12, 16]. 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), yet we implicitly condition on the fact that participants entered into the study ($S = 1$).

To illustrate the problem of selection bias, consider the following 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 implicit condition on $S = 1$.

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

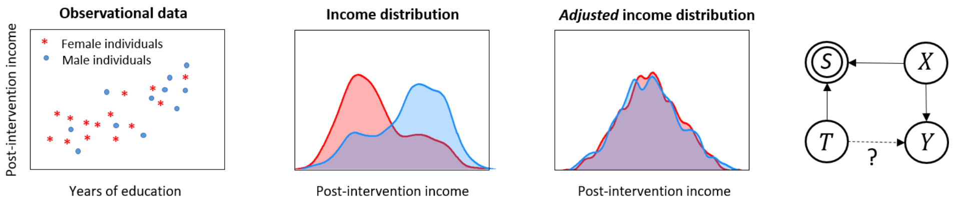

Figure 1: **The influence of selection bias**. The left panel plots a sample from the observed data, the middle panel shows the observed post-intervention income density for males and females, while the right panel shows the income distribution obtained by adjusting for education levels (we partition into homogeneous groups before aggregating their densities, see a description of the problem in the introduction and details of the data generating mechanism in Supplement B). In this case, a conventional two-sample test rejects the hypothesis of equal post-intervention income in male and female populations, while our proposed test fails to reject.

This problem has attracted recent interest in the causal inference literature [17]. [1] gave graphical conditions under which the causal effect may be recovered from data with selection bias, and can also be related to the problem of transportability which deals with transferring causal information from one environment to another, in which only passive observations can be collected [18]. 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 cause treatments and outcomes [25, 11, 27]. In epidemiology and econometrics, versions of this problem are also widely studied. The prevailing method is to adopt a model-based approach to treatment effect estimation, for example removing bias with knowledge of the probability of selection given treatment [19] or a probabilistic model of the selection mechanism [9].

The objective in all of the above is to give a *point estimate* of the causal effect. Much less is known however on the *significance* of treatment effects. Many empirical studies, especially those investigating treatments and effects from finite samples, *require* a notion of statistical significance to assess treatment outcomes. Current tests for the significance of treatment effects lack the flexibility of treatment effect estimation methods. 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 [4, 5] - both narrow summaries of outcome distributions. Researchers may be interested in the significance of effects beyond the average or conditional average - instead towards differences in the *whole* distribution of outcomes - and without relying on correctly specifying the model that links covariates, treatments and outcomes.

Contributions. This paper bridges this gap. We develop a non-parametric test for differences in distribution of treatment effect in two samples. Our proposal is a generalization of two sample tests based on maximum mean discrepancies [7, 3, 10, 26, 2] between probability distributions that incorporate importance sampling techniques to adjust for distributional shift in covariates. Conventional two-sample tests are recovered as a special case of our formulation that extends the realm of application of two-sample tests to decision-making with confounded data: a scenario where conventional two-sample tests fail. The technical challenge is that adjustments made for differences in the marginal confounding distributions between two samples are data-dependent, and therefore void existing asymptotic guarantees of tests based on the maximum mean discrepancy. We derive novel asymptotic distributions for the proposed test under conventional conditions on the causal structure between variables, and propose approximations with finite data we demonstrate to be consistent.

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 $T = T(X) \in \{0, 1\}$ 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 [20]. We assume to have observed independent samples from and 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, which appeal to the potential values of an individual were $T = 0$ and $T = 1$ 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 and 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 [14]. 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 (RKHS) $\mathcal{H}_k$. 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) [7] and is particularly appealing because for certain choices of the kernel function k, the mean embedding can be shown to be injective [22]. 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 [7],

$$\text{MMD}^2 := \underset{y,y^\star \sim P_{Y^1}}{\mathbb{E}} k(y, y^\star) + \underset{y,y^\star \sim P_{Y^0}}{\mathbb{E}} k(y, y^\star) - 2 \underset{y \sim P_{Y^1}, y^\star \sim P_{Y^0}}{\mathbb{E}} k(y, y^\star) \tag{3}$$

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 necessary 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,

$$\mathbb{E}\left(\frac{Y}{e(X)} \;\middle|\; T = 1\right) = \mathbb{E}\left(\frac{T \cdot Y^1}{e(X)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{T \cdot Y^1}{e(X)} \;\middle|\; X\right)\right) = \mathbb{E}\left(Y^1\right) \tag{4}$$

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,

$$\text{WMMD}^2 := \underset{P_{XY|T=1}}{\mathbb{E}} w(x)w(x^\star)k(y,y^\star) + \underset{P_{Y|T=0}}{\mathbb{E}} k(y,y^\star) - 2 \underset{x,y\sim P_{XY|T=1}, y^\star \sim P_{Y|T=0}}{\mathbb{E}} w(x)k(y,y^\star) \tag{5}$$

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 weights $w(x) > 0$ for all $x \in \mathcal{X}$, WMMD $= 0$ if and only if $P_{Y_1} = P_{Y^0}$.*

Proof. All proofs are given in the Supplementary material.

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

$$\widehat{\text{WMMD}}^2 := \sum_{i\neq j: t_i=t_j=1} w(x_i)w(x_j)k(y_i,y_j) + \sum_{i\neq j: t_i=t_j=0} k(y_i,y_j) - 2\sum_{i,j: t_i=1, t_j=0} w(x_i)k(y_i,y_j)$$

where the (y_i, t_i, x_i) are realizations of the random variables (Y, T, X). 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 $W \in \mathbb{R}^{n\times n}$ ($W_{ij} = w(x_i)w(x_j)$) be approximately diagonalizable (made precise in Supplement A.2). Then, the following statements hold,*

1. *Under $\mathcal{H}_0$, the asymptotic distribution of $\widehat{\text{WMMD}}^2$ is given by a mixture of independent χ^2 random variables. We provide the exact terms in Supplement A.2.*
2. *Under $\mathcal{H}_1$, $n^{1/2}\left(\widehat{\text{WMMD}}^2 - \text{WMMD}^2\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right)$*

We have used $\xrightarrow{d}$ to denote convergence in distribution. See Supplement A.2 for concrete expressions of all terms involved and a proof that relies on an approximate eigen-decomposition of the weight matrix and involves large-sample distributional approximations of quadratic forms and $U-$statistics.

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 [23] 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 [8].

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,

$$\hat{w} := \underset{0<w<B}{\operatorname{argmin}} \left|\left|\mathbb{E}_{P_{X|T=0}} w(x)\phi(x) - E_{P_{X|T=1}}\phi(x)\right|\right|_{\mathcal{H}_K} \tag{6}$$

This problem is convex. For injective mappings, minimizing (6) 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. In our implementation 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 (6) in the computation of $\widehat{\mathrm{WMMD}}$ and of the asymptotic null distribution still yields a consistent test.

Theorem 2 (Consistency of $\widehat{\mathrm{WMMD}}$). *Let $\hat{w}(x)$ be the empirical density ratio estimates of $w(x)$ - the underlying population 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 Supplement A.3, 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 $\widehat{\mathrm{WMMD}}$ 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 $\widehat{\mathrm{WMMD}}$ for the case where weights are upper-bounded by some maximum value.

Theorem 3 (Large deviation bound of $\widehat{\mathrm{WMMD}}$). *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,*

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

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 - 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.

4 Relationships to Testing in Regression Models and Other Tests

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 [21] for the outcome y,

$$Y_i = \mu + Z_i u_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

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 $\sigma_u^2 = 0$ [6, 15]. 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 [6]. The statistic S therefore has a high value whenever the terms of the matrix $K = (k_{ij})$ and the matrix $\tilde{Y}\tilde{Y}^T = (\tilde{Y}_i \tilde{Y}_j)$ are correlated. Now consider the case $n = m$ and write $y_{i,1} = y_i$ such that $t_i = 1$, and analogously for $y_{j,0}$, $i, j = 1, ..., n$. 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 $k(a, b)$.

Related work. As the example above shows, regression methods can be accommodated for treatment effect significance testing. One related approach are 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 [24]. Another approach is to replace the function in (7) with a non-parametric alternative, for example using power series as basis functions, for example as proposed in [4]. However, in these cases, the hypothesis being tested tends to be restricted to average effects. One extension to conventional 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 [26], proposed initially as more efficient alternatives to conventional tests. In our experiments, we implement non-parametric versions of each one of these, see Supplement C for details.

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. 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 [26], and finally the unweighted (conventional) MMD test [7]. For kernel-based tests, since no closed-form quantiles, in each trial we use 400 random permutations to approximate the null.

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. We consider the following data distributions for two samples of data ($i = 0, 1$) that exhibit a spurious dependence between random variables such as might occur due to selection bias,

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

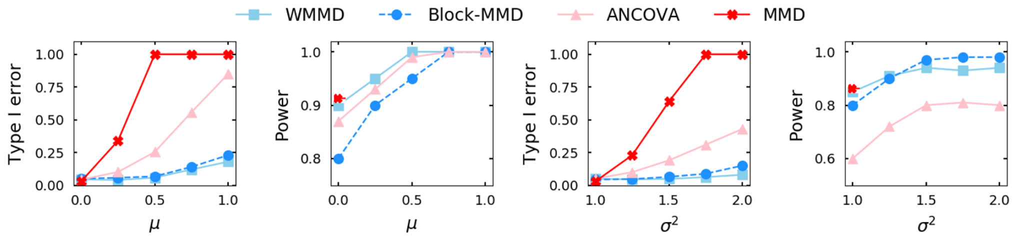

Figure 2: Type I error (lower better) and Power (higher better) of all tests on synthetic experiments. The WMMD has simultaneously best control of type I errors and highest power.

μ and σ^2 determine selection bias, i.e. the extent of the dependence between X and T. For instance $\mu > 0$ will result in more density in regions of the support of X. 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), 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: Difference in means μ (with $\sigma^2 = 1$) and difference in variances σ^2 (with $\mu = 0$) of confounding variables. In each case the dimensionality of X and Y is 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 $g_1(x) = [\sin(x_1), x_2, ..., x_{20}]$. The latter is a challenging problem as only the first dimension varies.

We observe in Figure 2 that WMMD maintains 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. Notice 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. Power increases with confounder distributional shift, which is expected 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 displays higher or competitive power. We make an important comparison also in the two power experiments in the absence of 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.

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 confounding 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" - the MMD computed from data with no unobserved confounding). With increasing confounding, we see in the leftmost 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 - see this with 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.

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

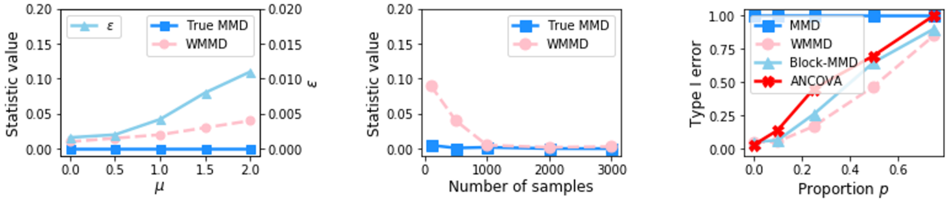

Figure 3: The two leftmost panels show the approximation error of the WMMD with increasing confounding and increasing sample size - see details in section 5.1.2. The rightmost panel show type I errors in the presence of unobserved confounders - see details in section 5.1.3.

specifications considered above. To do so, we remove from the observed data a proportion p of variables X. The results 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 (a criticism however that applies to all most treatment effect estimation algorithms). 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). In some cases we may rule out both of the above 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, 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 in the mid-1970s in the U.S. [13]. The outcome of interest is earnings two years after the end of the employment program. Our null hypothesis is no difference in earningsdue to the program. Posterior earnings in treated and control populations are not directly comparable because the populations differ systematically in their education level, prior earnings, age, etc., all plausible confounders. The data contains 614 individuals, 185 of whom were included in the employment program. With real data, the ground truth relationship between two populations is unknown. To evaluate tests for this problem, 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 type I error be appropriately controlled. These results as a functin of the proportion p of observations removed (increasing bias) is given in Table 1. On the original data, all tests returned significant earning difference.

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 as a function of bias p.

6 Conclusions

We have proposed a test statistic for the two-sample problem that expands the toolkit of statisticians to make inference for treatment effects on biased data. Bias in the sample selection mechanism creates distributional shift which leads to bias in the treatment effect if unaccounted for, a context which is unexplored in hypothesis testing. Our proposal is a generalization of the MMD to adjust for this bias, we have demonstrated our test to be consistent for the problem of interest, derived its asymptotic distribution and derived large deviation bounds. In empirical comparisons, we have shown our test to be more powerful than existing alternatives while controlling approximately for type I error.

7 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

[1] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

[2] Alexis Bellot and Mihaela van der Schaar. Kernel hypothesis testing with set-valued data. *arXiv preprint arXiv:1907.04081*, 2019.

[3] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

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

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

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

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

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

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

[10] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271, 2017.

[11] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[12] David S Kirk, Geoffrey C Barnes, Jordan M Hyatt, and Brook W Kearley. The impact of residential change and housing stability on recidivism: pilot results from the maryland opportunities through vouchers experiment (move). *Journal of experimental criminology*, 14(2):213–226, 2018.

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

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

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

[16] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM, 2019.

[17] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[18] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 540–547. IEEE, 2011.

[19] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

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

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

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

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

[24] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 7. Pearson Boston, MA, 2019.

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

[26] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763, 2013.

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