Biometrika (2025), **xx**, x, *p*. 1
Printed in Great Britain

General Frameworks for Conditional Two-Sample Testing

By Seongchan Lee*
Department of Statistics and Data Science, Yonsei University, Seoul, South Korea.
statchan1106@yonsei.ac.kr

Suman Cha*
Department of Statistics and Data Science, Yonsei University, Seoul, South Korea.
oldrain123@yonsei.ac.kr

Ilmun Kim
Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, Seoul, South Korea.
ilmun@yonsei.ac.kr

* these authors contributed equally to this work.

Summary

We study the problem of conditional two-sample testing, which aims to determine whether two populations have the same distribution after accounting for confounding factors. This problem commonly arises in various applications, such as domain adaptation and algorithmic fairness, where comparing two groups is essential while controlling for confounding variables. We begin by establishing a hardness result for conditional two-sample testing, demonstrating that no valid test can have significant power against any single alternative without proper assumptions. We then introduce two general frameworks that implicitly or explicitly target specific classes of distributions for their validity and power. Our first framework allows us to convert any conditional independence test into a conditional two-sample test in a black-box manner, while preserving the asymptotic properties of the original conditional independence test. The second framework transforms the problem into comparing marginal distributions with estimated density ratios, which allows us to leverage existing methods for marginal two-sample testing. We demonstrate this idea in a concrete manner with classification and kernel-based methods. Finally, simulation studies are conducted to illustrate the proposed frameworks in finite-sample scenarios.

Some key words: Conditional independence testing; Covariate shift; Density ratio estimation

1. Introduction

This paper addresses the problem of testing for equivalence between two conditional distributions, namely conditional two-sample testing. Statistical methods for this problem have important applications across diverse fields such as domain adaptation and algorithmic fairness. In domain adaptation, for instance, this methodology can serve as a formal framework to validate the covariate shift assumption, where the conditional distribution of Y given X remains unchanged, while the marginal distributions of X may differ. By confirming this assumption, practitioners can effectively re-weight the training data according to the marginal density ratio regarding X, which potentially leads to improved predictive performance and better adaptation to new domains (Shimodaira, 2000; Sugiyama et al., 2007a,b). Moreover, in algorithmic fairness, conditional two-sample testing plays a role in detecting and mitigating biases. In particular, it helps identify whether a certain machine learning model unfairly favors or disadvantages specific groups based on demographic characteristics such as age, gender, or ethnicity (Hardt et al., 2016; Barocas et al., 2023). Conditional two-sample testing also finds applications beyond machine learning. In genomics, for example, scientists seek to identify differences in genetic distributions conditional on various factors such as disease status and environmental exposures (Virolainen et al., 2022; Wu et al., 2023). This methodology aids scientists in understanding the genetic basis of diseases and in developing strategies for personalized medicine by providing a rigorous framework for comparing conditional distributions.

1.1. *Problem Setup*

With the practical motivation in mind, we now formally set up the problem. Given $n_1, n_2 \in \mathbb{N}$, suppose we observe two mutually independent samples

$$\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(1)} \quad \text{and} \quad \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(2)},$$

where $P_{XY}^{(1)}$ and $P_{XY}^{(2)}$ are joint distributions supported on some generic product space $\mathcal{X} \times \mathcal{Y}$. Let $P_{Y|X}^{(1)}$ and $P_{Y|X}^{(2)}$ denote the conditional distributions of $Y^{(1)} \mid X^{(1)}$ and $Y^{(2)} \mid X^{(2)}$, respectively. Similarly, let $P_X^{(1)}$ and $P_X^{(2)}$ denote the marginal distributions of $X^{(1)}$ and $X^{(2)}$, respectively. Given these two samples, our goal is to test the equality of two conditional distributions

$$H_0 : P_X^{(1)}\left\{P_{Y|X}^{(1)}(\cdot \mid X) = P_{Y|X}^{(2)}(\cdot \mid X)\right\} = 1 \;\text{ vs }\; H_1 : P_X^{(1)}\left\{P_{Y|X}^{(1)}(\cdot \mid X) \neq P_{Y|X}^{(2)}(\cdot \mid X)\right\} > 0, \quad (1)$$

where $P_X^{(j)}(\cdot \mid x)$ denotes the conditional distribution of $Y^{(j)}$ given $X^{(j)} = x$ for $j = 1, 2$. In other words, we are interested in determining whether two populations have the same distribution after controlling for potential confounding variables. Throughout this paper, we assume that $P_X^{(1)}$ and $P_X^{(2)}$ have the same support, satisfying $P_X^{(1)} \ll P_X^{(2)}$ and $P_X^{(2)} \ll P_X^{(1)}$ where the symbol $\ll$ denotes absolute continuity. Since $P_X^{(1)}$ and $P_X^{(2)}$ have the same support, the above hypotheses (1) for conditional two-sample testing can be equivalently defined using $P_X^{(2)}$ instead of $P_X^{(1)}$.

As pointed out by Boeken & Mooij (2021) and Yan & Zhang (2022), conditional two-sample testing is closely connected to conditional independence testing. To illustrate this connection, we introduce a binary variable $Z \in \{1, 2\}$, and see that the conditional independence between Y and Z given X is equivalently expressed as

$$Y \perp\!\!\!\perp Z \mid X \iff Y \mid X, Z = 1 \overset{d}{=} Y \mid X, Z = 2, \quad (2)$$

where the symbol $\stackrel{d}{=}$ denotes equality in distribution. This equivalence enables us to convert the problem of conditional two-sample testing to that of conditional independence testing based on the datasets $\{(Y_i, X_i) : Z_i = 1\}$ and $\{(Y_i, X_i) : Z_i = 2\}$. Consequently, we can leverage various existing methods for conditional independence testing to tackle conditional two-sample testing. However, prior work has not rigorously explored this approach, and indeed Yan & Zhang (2022) claim that it is not a sensible approach as the variable Z in the conditional two-sample problem is deterministic. Specifically, letting $n = n_1 + n_2$, $\sum_{i=1}^{n} \mathbb{1}(Z_i = 1)$ and $\sum_{i=1}^{n} \mathbb{1}(Z_i = 2)$ correspond to the sample sizes for two populations (i.e., n_1 and n_2), which are fixed in advance for the conditional two-sample problem. Therefore, a gap remains in rigorously connecting these seemingly similar, yet distinct, problems.

1.2. *An Overview of Our Results*

In this work, we make several contributions to the field of conditional two-sample testing. First, we reaffirm that comparing conditional distributions is intrinsically more difficult than comparing marginal distributions. For marginal two-sample testing, one can design permutation tests that control the type I error, while being powerful against certain alternatives (e.g., Kim et al., 2022a). However, we show that this is not the case for conditional two-sample testing. Our result (Theorem 1) proves that any valid conditional two-sample test has power at most equal to its size against any single alternative if the type of a conditional random vector is continuous. This is reminiscent of the negative result for conditional independence testing proved in Shah & Peters (2020). It is worth highlighting, however, that their negative result does not directly imply our Theorem 1. The proof of Shah & Peters (2020) relies on the assumption that the data $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$ are i.i.d., which does not hold in our setup as $\sum_{i=1}^{n} \mathbb{1}(Z_i = 1)$ and $\sum_{i=1}^{n} \mathbb{1}(Z_i = 2)$ are deterministic numbers. We handle this distinction through a concentration argument and show that conditional two-sample testing is as difficult as conditional independence testing. This negative result naturally motivates additional assumptions that make the problem feasible.

Our next contribution is to introduce two general frameworks for conditional two-sample testing. The first framework effectively addresses the issue pointed out by Yan & Zhang (2022). In particular, we develop a generic method that converts any conditional independence test into a conditional two-sample test. This general method directly transfers the asymptotic properties of a conditional independence test computed using $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} \stackrel{\text{i.i.d.}}{\sim} P_{XYZ}$ to the setting of conditional two-sample testing (Theorem 2). At the heart of this approach is the concentration property of a Binomial random variable to its mean, which facilitates the effective construction of i.i.d. samples drawn from P_{XYZ} (see Algorithm 1). This development paves a way to leverage any existing methods for conditional independence testing in the literature, thereby expanding the range of tools available to practitioners for conducting two-sample tests.

The second framework that we introduce is based on density ratio estimation. To elaborate, let us assume that $P_X^{(1)}$ and $P_X^{(2)}$ have density functions $f_X^{(1)}$ and $f_X^{(2)}$ with respect to some base measure, and similarly $P_{Y|X}^{(1)}(\cdot \mid x)$ and $P_{Y|X}^{(2)}(\cdot \mid x)$ have density functions $f_{Y|X}^{(1)}(\cdot \mid x)$ and $f_{Y|X}^{(2)}(\cdot \mid x)$, respectively. Then for all $x, y \in \mathcal{X} \times \mathcal{Y}$, we have the identity:

$$f_{Y|X}^{(1)}(y \mid x) = f_{Y|X}^{(2)}(y \mid x) \iff f_{YX}^{(1)}(y, x) = f_{YX}(y, x), \tag{3}$$

where $f_{YX}(y,x) = f^{(2)}_{YX}(y,x)\, f^{(1)}_X(x)/f^{(2)}_X(x)$ and $f^{(1)}_{YX}$ is the joint density function of $(Y^{(1)}, X^{(1)})$ such that $f^{(1)}_{YX}(y,x) = f^{(1)}_{Y|X}(y \mid x) f^{(1)}_X(x)$, and $f^{(2)}_{YX}$ is similarly defined for $(Y^{(2)}, X^{(2)})$. The above equivalence (3) allows us to transform the problem of testing for conditional distributions into the one that compares marginal distributions with densities $f^{(1)}_{YX}$ and f_{YX}. The latter problem has been extensively studied with various methods, ranging from classical approaches such as Hotelling's test to modern methods such as kernel maximum mean discrepancy (Gretton et al., 2012; Liu et al., 2020; Schrab et al., 2023) and machine learning-based approaches (e.g., Lopez-Paz & Oquab, 2017; Kim et al., 2019, 2021; Hediger et al., 2022). The issue, however, is that we do not observe samples from f_{YX} but from $f^{(2)}_{YX}$. Therefore, the success of this framework relies on how accurate one can estimate the density ratio $r_X(x) = f^{(1)}_X(x)/f^{(2)}_X(x)$ and incorporate it into the procedure to fill the gap between f_{YX} and $f^{(2)}_{YX}$. We demonstrate this methodology focusing on a classification-based test in Section 4.1 and a kernel-based test in Section 4.2.

1.3. *Literature Review*

As mentioned earlier, conditional two-sample testing has a wide range of applications in various fields, including machine learning, genetics and economics, where it is important to compare two samples controlling for confounding variables. Despite its broad range of applications and significance, there has been limited research dedicated to tackling this fundamental problem. Similar problems, on the other hand, have been explored in the literature such as testing for the equality of conditional moments (Hall & Hart, 1990; Kulasekera, 1995; Kulasekera & Wang, 1997; Fan & Lin, 1998; Neumeyer & Dette, 2003; Pardo-Fernández et al., 2015) and goodness-of-fit testing for pre-specified conditional distributions (Andrews, 1997; Zheng, 2000; Fan et al., 2006). These methods aim to facilitate the comparison of specific aspects of a distribution such as the conditional mean or second moments, rather than the entire distribution. Our research, however, is centered on nonparametric comparisons of two conditional distributions. This approach is of great importance as it enables a more comprehensive comparison of distributions, capturing differences that may not be evident when only specific moments or pre-specified models are compared.

It is only in recent years that conditional two-sample testing has gained attention, with several novel methods being proposed. Yan & Zhang (2022), for instance, proposed a method that extends unconditional energy distance to its conditional counterpart. They demonstrated that many key properties of the unconditional energy distance are retained in the conditional version. Moreover, they proposed a bootstrap procedure to calibrate their test statistic. To the best of our knowledge, however, the validity of their test remains unexplored, and the $O(n^4)$ time complexity of their algorithm poses a bottleneck to its practical application.

As another example, Hu & Lei (2024) built on the idea of conformal prediction and introduced a nonparametric conditional two-sample test using a weighted rank-sum statistic. This approach involves estimating both marginal and conditional density ratios, and the validity of their method depends on the quality of these ratio estimators. As explained in Example 4, their test statistic can be viewed as an example of our general framework based on density ratio estimation. A more recent work by Chen & Lei (2024) extended the idea of Hu & Lei (2024), leveraging Neyman orthogonality to reduce the first-order bias for the asymptotic normality. As another closely related work, Chatterjee et al. (2024) introduced a kernel-based conditional two-sample test using nearest

neighbors. They considered the setting where a random sample $\{(X_i, Y_i^{(1)}, Y_i^{(2)})\}_{i=1}^n$ is generated from a joint distribution, i.e., the response variables $Y^{(1)}$ and $Y^{(2)}$ are conditioned on the same set of covariates X. This setting is notably different from that considered in the prior work (Yan & Zhang, 2022; Hu & Lei, 2024; Chen & Lei, 2024) as well as in our study, which consider potentially different covariates. Hence, the methods proposed by Chatterjee et al. (2024) are not directly comparable to ours.

As explained before, the first framework that we propose can be constructed based on essentially any conditional independence tests from the literature. The problem of testing for conditional independence has been extensively studied, resulting in a variety of methods to handle different scenarios and challenges. Shah & Peters (2020) proposed the Generalized Covariance Measure (GCM) whose validity depends on the performance of regression methods. Recent improvements to this method include the strategies such as weighting (Scheidegger et al., 2022) and applying GCM to a projected random vector (Lundborg et al., 2022; Chakraborty et al., 2024). Other notable methodologies for conditional independence testing include kernel-based tests (Zhang et al., 2011; Doran et al., 2014; Strobl et al., 2019; Pogodin et al., 2024), binning-based tests (Neykov et al., 2021; Kim et al., 2022b; Neykov et al., 2023), regression-based tests (Dai et al., 2022; Williamson et al., 2023) and tests under the model-X framework (Candes et al., 2018; Berrett et al., 2020; Liu et al., 2022; Tansey et al., 2022). Our method can leverage these developments to effectively solve the problem of conditional two-sample testing.

Our second framework can benefit from extensive research done on density ratio estimation in the literature. A straightforward way of estimating density ratio is to first estimate individual density functions, and take their ratio as an estimate. However, this method tends to become unstable, especially in high-dimensional settings. To overcome this issue, Sugiyama et al. (2007b) and Tsuboi et al. (2009) developed methods that directly estimate density ratio without involving density estimation. Kanamori et al. (2010) compared different methods of density ratio estimation, and discussed their theoretical properties. Kanamori et al. (2009) reformulated the problem as a least-squares problem to provide a closed-form solution, whereas Liu et al. (2017) proposed trimmed density ratio estimation to improve stability and robustness by trimming extreme values. More recent advancements in density ratio estimation include Choi et al. (2021); Rhodes et al. (2020); Choi et al. (2022). As explained in Section 4, our approach uses density ratio estimation to deal with discrepancies between $f_{YX}^{(1)}$ and f_{YX}, and transforms the problem of comparing conditional distributions into that of comparing marginal distributions.

1.4. *Organization*

The rest of this paper is organized as follows. We begin with a hardness result for conditional two-sample testing in Section 2, which shows that no test can have power greater than its size against any alternative without additional assumptions. Section 3 presents our framework that converts tests for conditional independence into those for the equality of conditional distributions. Section 4 introduces another framework based on density ratio estimation. Numerical results illustrating the finite-sample performance of our methods are presented in Section 5, followed by the conclusion in Section 6. The proofs of the results can be found in the supplementary material.

2. Hardness Result

Before introducing our frameworks, we present a fundamental hardness result for conditional two-sample testing. Specifically, for a continuous random vector X, our result demonstrates that any valid conditional two-sample test has no power against any alternative. This finding parallels the negative result established by Shah & Peters (2020) for conditional independence testing, and our proof builds crucially on their work. Given the connection established in (2), one might argue that their negative result directly applies to the two-sample problem. However, additional effort is required to make this connection concrete since the sample sizes n_1 and n_2 are deterministic in our setting, which violates the i.i.d. assumption required in Shah & Peters (2020).

To state the result, let (X, Y, Z) be a random vector in $\mathbb{R}^{d_X+d_Y} \times \{1, 2\}$, and $\mathcal{E}$ be the set of all distributions such that X and Y have marginal distributions, which are absolutely continuous with respect to the Lebesgue measure. In addition, Z has a discrete distribution supported on $\{1, 2\}$ with $\mathrm{pr}(Z = 1) = \lambda_n \in (0, 1)$ where $\lambda_n = n_1/n$. Let $\mathcal{P}_0 \subset \mathcal{E}$ be the set of null distributions for conditional independence, i.e., distributions such that $Y \perp\!\!\!\perp Z \mid X$, and let $\mathcal{P}_1 = \mathcal{E} \setminus \mathcal{P}_0$ be the set of alternative distributions. For $M \in (0, \infty]$, let $\mathcal{E}_M \subseteq \mathcal{E}$ be the subset of all distributions where X and Y are supported within an ℓ_∞ ball of radius M. Set $\mathcal{P}_{0,M} = \mathcal{P}_0 \cap \mathcal{E}_M$ and $\mathcal{P}_{1,M} = \mathcal{P}_1 \cap \mathcal{E}_M$. Further denoting $N_1 = \sum_{i=1}^n \mathbb{1}(Z_i = 1)$ and $N_2 = \sum_{i=1}^n \mathbb{1}(Z_i = 2)$, the following theorem shows that no valid test ϕ for conditional two-sample testing has power greater than its size.

Theorem 1. *Let $n_1, n_2 \in \mathbb{N}$ with $n = n_1 + n_2$, $\alpha \in (0, 1)$ and $M \in (0, \infty]$. For $\{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XYZ} = P$, consider a test $\phi : \{(X_i, Y_i, Z_i)\}_{i=1}^n \mapsto \{0, 1\}$. Suppose that ϕ controls the type I error at level α conditional on $N_1 = n_1$ and $N_2 = n_2$ as*

$$\sup_{P \in \mathcal{P}_{0,M}} E_P(\phi \mid N_1 = n_1, N_2 = n_2) \leq \alpha.$$

Then the power of ϕ conditional on $N_1 = n_1$ and $N_2 = n_2$ is at most α for any $P \in \mathcal{P}_{1,M}$ as

$$E_P(\phi \mid N_1 = n_1, N_2 = n_2) \leq \alpha.$$

Remark. We note that the type I error and the power of ϕ conditional on $N_1 = n_1$ and $N_2 = n_2$ in Theorem 1 are equivalent to those computed under the two-sample setting with deterministic sample sizes. In other words, the testing errors for conditional two-sample testing can be understood as the testing errors for conditional independence testing conditional on $N_1 = n_1$ and $N_2 = n_2$; thereby Theorem 1 implying the negative result for conditional two-sample testing. We also note that Theorem 1 only focuses on non-randomized tests for simplicity of presentation, but our proof also holds for randomized tests. The proof of Theorem 1 is provided in the supplementary material.

Theorem 1 clearly explains that it is necessary to impose additional assumptions (e.g., smoothness for distributions) in order to make the conditional two-sample problem feasible. In the next two sections, we explore two general frameworks, which implicitly or explicitly incorporate reasonable assumptions to address this problem. The first framework utilizes any conditional independence test and considers scenarios where this test performs well for verifying conditional independence. Conversely, the second framework assumes that the marginal density ratio r_X is well-behaved and can be estimated with high accuracy.

3. Approach via Conditional Independence Testing

In this section, we introduce our first framework that converts a conditional independence test to a conditional two-sample test, while maintaining the same asymptotic guarantees. The key idea is to construct a dataset $\mathcal{D}_{\tilde{n}}$ consisting of i.i.d. random vectors (Y, Z, X) of size $\tilde{n}$ based on the given two samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$. To achieve this, letting $n = n_1 + n_2$, we first draw a random variable $\tilde{n}_1$ from Binomial$(\tilde{n}, n_1/n)$ where $\tilde{n}$ is set to be smaller than n and $\tilde{n}/n \to 1$. Since a Binomial random variable is highly concentrated around its mean, we can guarantee that $\tilde{n}_1 \le n_1$ and $\tilde{n}_2 := \tilde{n} - \tilde{n}_1 \le n_2$ with high probability. If a bad event happens where either $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$, making the construction of $\mathcal{D}_{\tilde{n}}$ infeasible, we simply accept the null hypothesis. This slightly inflates the type II error in finite-sample scenarios, but it is asymptotically negligible. A similar idea has been utilized in Neykov et al. (2021) in a different context to eliminate Poissonization for conditional independence testing.

Having constructed $\mathcal{D}_{\tilde{n}}$ consisting of i.i.d. random samples drawn from the joint distribution of (Y, Z, X), we can now implement a conditional independence test based on $\mathcal{D}_{\tilde{n}}$, while retaining the same theoretical guarantees for conditional two-sample testing. Algorithm 1 summarizes this procedure, and the following theorem formally establishes its theoretical guarantees.

Theorem 2. *Consider a class of distributions $\mathcal{P}$ of (X, Y, Z) where Z takes a value among $\{1, 2\}$ with probability n_1/n and n_2/n, respectively, and let $\{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P \in \mathcal{P}$. For disjoint subclasses $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1 \subset \mathcal{P}$ and $\alpha \in (0, 1)$, assume that a test $\phi : \{(X_i, Y_i, Z_i)\}_{i=1}^n \mapsto \{0, 1\}$ satisfies*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} E_P(\phi) \le \alpha \quad \text{and} \quad \lim_{n\to\infty} \sup_{P\in\mathcal{P}_1} E_P(1-\phi) = 0.$$

Denote the output of Algorithm 1 with $\varepsilon = o(1)$ as $\tilde{\phi} \in \{0, 1\}$ where $\tilde{\phi} = 1$ if and only if H_0 is rejected. Then it holds that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} E_P(\tilde{\phi} \mid N_1 = n_1, N_2 = n_2) \le \alpha \quad \text{and} \quad \lim_{n\to\infty} \sup_{P\in\mathcal{P}_1} E_P(1-\tilde{\phi} \mid N_1 = n_1, N_2 = n_2) = 0,$$

where we recall $N_1 = \sum_{i=1}^n \mathbb{1}(Z_i = 1)$ and $N_2 = \sum_{i=1}^n \mathbb{1}(Z_i = 2)$.

Algorithm 1. Converting a Conditional Independence Test into a Conditional Two-Sample Test

Input:

Data $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$ of sizes n_1 and n_2,
a conditional independence test ϕ for $H_0 : Y \perp\!\!\!\perp Z \mid X$ of (asymptotic) size $\alpha \in (0, 1)$,
adjustment parameter $\varepsilon \in (0, 1)$.

Draw $\tilde{n}_1 \sim$ Binomial$(\tilde{n}, n_1/n)$ where $\tilde{n} = k^* n$ and

$$k^* = 1 - \frac{3\log(\varepsilon)}{2n_1} - \left\{\left(1 - \frac{3\log(\varepsilon)}{2n_1}\right)^2 - 1\right\}^{1/2} \text{ and set } \tilde{n}_2 = \tilde{n} - \tilde{n}_1.$$

If $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$ **then** Accept H_0.

Else

Set $\mathcal{D}_{\tilde{n}} = \{(X_i^{(1)}, Y_i^{(1)}, Z_i = 1)\}_{i=1}^{\tilde{n}_1} \cup \{(X_i^{(2)}, Y_i^{(2)}, Z_i = 2)\}_{i=1}^{\tilde{n}_2}$.
Run ϕ using $\mathcal{D}_{\tilde{n}}$ at level α. Denote the resulting test function as $\phi_{\tilde{n}}$.
If $\phi_{\tilde{n}} = 1$ **then** Reject H_0 **else** Accept H_0.

The proof of Theorem 2 is provided in supplementary material. Our analysis in Theorem 2 is not limited to conditional two-sample testing, and it can be applied to marginal two-sample testing as well. Indeed, the problem of conditional two-sample testing becomes equivalent to the unconditional counterpart when X is degenerate (e.g., $X = 0$ with probability one). Thus, our algorithm serves as a generic method to convert unconditional independence tests to unconditional two-sample tests as well. We also mention that a specific form of k^* in Algorithm 1 is derived from the multiplicative Chernoff bound for a Binomial random variable (Lemma 1 of supplementary material), which can be refined by numerically computing the tail probability of a Binomial random variable.

Despite its generality, one obvious drawback of Algorithm 1 is that it does not take the datasets $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=\tilde{n}_1+1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=\tilde{n}_2+1}^{n_2}$ into account in the procedure when $\tilde{n}_1 < n_1$ and $\tilde{n}_2 < n_2$. It can be seen that the expected number of discarded samples, i.e., $E(n - \tilde{n}_1 - \tilde{n}_2)$, is $O\big(\{n \log(1/\varepsilon)\}^{1/2}\big)$. This loss might degrade the performance in small-sample size regimes, but it can be negligible when n is sufficiently large and ε decreases slowly (see supplementary material for empirical support). Nevertheless, when a test statistic is sufficiently stable, the conclusion of Theorem 2 may hold without further modification of ϕ, meaning the conditional testing errors of ϕ are asymptotically equivalent to its marginal errors.

To illustrate this, we build upon the coupling argument presented by Chung & Romano (2013). First, draw $\bar{n}_1 \sim \text{Binomial}(n, n_1/n)$ and set $\bar{n}_2 = n - \bar{n}_1$. If $\bar{n}_1 > n_1$, draw $\bar{n}_1 - n_1$ additional samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=n_1+1}^{\bar{n}_1}$ from $P_{XY}^{(1)}$. Otherwise, draw $\bar{n}_2 - n_2$ additional samples $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=n_2+1}^{\bar{n}_2}$ from $P_{XY}^{(2)}$. In either case, set $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^{n} = \{X_i^{(1)}, Y_i^{(1)}, 1\}_{i=1}^{\bar{n}_1} \cup \{X_i^{(2)}, Y_i^{(2)}, 2\}_{i=1}^{\bar{n}_2}$, which can be viewed as i.i.d. draws from a joint distribution of (X, Y, Z), after randomly permuting indices. When this newly constructed dataset is compared with the original dataset $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} = \{X_i^{(1)}, Y_i^{(1)}, 1\}_{i=1}^{n_1} \cup \{X_i^{(2)}, Y_i^{(2)}, 2\}_{i=1}^{n_2}$, there are $|\bar{n}_1 - n_1|$ distinct data points with the expectation $E[|\bar{n}_1 - n_1|] \le \sqrt{n/4}$. This suggests that if a test statistic T is asymptotically invariant to $n^{1/2}$-data perturbations, the asymptotic behavior of T remains consistent across both the original and the newly constructed datasets. However, the stability of T needs to be evaluated on a case-by-case basis. Below, we provide examples illustrating both stable and unstable cases.

Example 1 (Stable case). To simplify our presentation, consider a univariate case of $Y \in \mathbb{R}$, and assume $f(x) = E(Y \mid X = x)$ and $g(x) = E(Z \mid X = x)$ are known. Letting $R_i = \{Y_i - f(X_i)\}\{Z_i - g(X_i)\}$, the generalized covariance measure introduced by Shah & Peters (2020) is

$$T = \frac{\frac{1}{n^{1/2}} \sum_{i=1}^{n} R_i}{\left\{\frac{1}{n} \sum_{i=1}^{n} R_i^2 - \left(\frac{1}{n} \sum_{r=1}^{n} R_r\right)^2\right\}^{1/2}},$$

and let $\tilde{T}$ be similarly defined as T based on $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^{n}$. Focusing on the numerators of T and $\tilde{T}$, it can be seen that their difference is

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} R_i - \frac{1}{n^{1/2}} \sum_{i=1}^{n} \tilde{R}_i = \frac{1}{n^{1/2}} \sum_{i=n_1+1}^{\bar{n}_1} (R_i - \tilde{R}_i) \mathbb{1}(\bar{n}_1 > n_1) + \frac{1}{n^{1/2}} \sum_{i=\bar{n}_1+1}^{n_1} (R_i - \tilde{R}_i) \mathbb{1}(\bar{n}_1 \le n_1).$$

Under the null hypothesis, the expectation of the difference is zero and the variance is bounded above by $1/n^{1/2}$ up to a constant, provided that each Y_i has a finite second moment. Therefore,

the difference of the numerators is asymptotically negligible. We can show similarly that the difference of the denominators is also asymptotically negligible as detailed in supplementary material. Putting things together concludes that T and $\tilde{T}$ are asymptotically equivalent.

In the above example, we assumed that the conditional expectations f and g are known. In practice, f and g are estimated from the data, and the generalized covariance measure can become highly unstable when the estimators of f and g are themselves unstable.

Example 2 (Unstable case). Consider an extreme case where the estimators $\hat{f}$ and $\hat{g}$ are defined as follows: $\hat{f}(X_i) = Y_i \mathbb{1}(\sum_{i=1}^n Z_i = n_1) + f(X_i)\mathbb{1}(\sum_{i=1}^n Z_i \neq n_1)$ and $\hat{g}(X_i) = Z_i \mathbb{1}(\sum_{i=1}^n Z_i = n_1) + g(X_i)\mathbb{1}(\sum_{i=1}^n Z_i \neq n_1)$. In this case, T is not well-defined as it takes the form $0/0$ deterministically. On the other hand, when n_1 and n_2 are well-balanced (e.g., $n_1/n = 1/2$), the probability of the event $\sum_{i=1}^n \tilde{Z}_i = n_1$ converges to zero as n increases. Under such condition and assuming suitable moment conditions, the test statistic $\tilde{T}$ based on $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$ can still converge to a Gaussian limit. This example illustrates that the limiting behavior of T and $\tilde{T}$ can differ significantly, which is attributed to the instability of the estimators $\hat{f}$ and $\hat{g}$.

The previous examples highlight the need for caution when converting a conditional independence test to a conditional two-sample test, and also justify our generic approach to converting a conditional independence test to a conditional two-sample test in Algorithm 1. The next section introduces another general framework for conditional two-sample testing based on density ratio estimation.

4. Approach via Density Ratio Estimation

In this section, we present our second framework, which transforms the problem of conditional two-sample testing into one that involves comparing marginal distributions via density ratio estimation. Concretely, we recall from (3) that the null hypothesis of equality of two conditional distributions holds if and only if $f_{YX}^{(1)} = f_{YX}$ where $f_{YX} = r_X f_{YX}^{(2)}$ and r_X is the density ratio defined in Section 1.2. A challenge when applying this approach is that we only have samples from $f_{YX}^{(2)}$, and not from f_{YX}, which makes it impossible to directly compare samples from $f_{YX}^{(1)}$ with those from f_{YX}. However, once the density ratio is known or accurately estimated, we can effectively correct the bias arising from the difference between $f_{YX}^{(2)}$ and f_{YX} in various test statistics, frequently used for marginal two-sample testing. To facilitate our discussion, we first assume that the density ratio r_X is known and provide a detailed analysis on how to deal with the unknown case by focusing on a few cases.

At the core of our idea is importance weighting (Kimura & Hino, 2024, for a survey), a technique that assigns different levels of importance to data points to correct biases and prioritize relevant data. For instance, suppose we would like to estimate the expectation of X under the distribution P with density p, while we only observe data $X_1, \ldots, X_n$ from another distribution Q with density q. Then by re-weighting data points using the density ratio p/q, we can obtain an unbiased estimator of the expectation under P as

$$\frac{1}{n}\sum_{i=1}^n \frac{p(X_i)}{q(X_i)} X_i \quad \text{such that} \quad E_Q\left\{\frac{1}{n}\sum_{i=1}^n \frac{p(X_i)}{q(X_i)} X_i\right\} = E_P(X).$$

This idea can be applied to a range of marginal two-sample test statistics as we demonstrate below. Throughout this section, we use the shorthand $V^{(1)} = (X^{(1)}, Y^{(1)})$ and $V^{(2)} = (X^{(2)}, Y^{(2)})$ to simplify the notation.

Example 3. **(Mean comparison)** We start with a simple case of comparing the mean of transformed samples. Given a feature map $\psi : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, one can consider

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\psi\big(V_i^{(1)}\big) - \frac{1}{n_2}\sum_{i=1}^{n_2} r_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big)$$

as a test statistic for the hypotheses in (1). The expectation of this statistic is equal to zero under the null hypothesis. Moreover, since the test statistic is simply a linear combination of independent random variables, it can be calibrated using the Gaussian approximation.

Example 4. **(Rank sum statistic)** Instead of comparing the mean, one can compare the stochastic order of two distributions using ranks. Specifically, given a feature map $\psi : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, a rank sum statistic based on the transformed samples can be computed as

$$\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_X\big(X_j^{(2)}\big)\mathbb{1}\big\{\psi\big(V_j^{(2)}\big) < \psi\big(V_i^{(1)}\big)\big\}.$$

Under the null hypothesis, and assuming no ties among transformed samples, it can be seen that the expectation is equal to zero. As in Hu & Lei (2024), the test statistic can be shown to be asymptotically Gaussian using the asymptotic theory of U-statistics under conditions. Therefore, the critical value can be determined based on this Gaussian approximation. The power, however, changes depending on ψ. Hu & Lei (2024) takes ψ as an estimate of $f^{(1)}_{Y|X}(\cdot \mid \cdot)/f^{(2)}_{Y|X}(\cdot \mid \cdot)$.

Example 5. **(Classifier-based approach)** Let $\mathcal{H}$ be a class of classifiers. Given a binary classifier $h : \mathcal{X} \times \mathcal{Y} \mapsto \{1, 2\}$ where $h \in \mathcal{H}$, $\ell : \mathbb{R} \times \{1, 2\} \mapsto \mathbb{R}$ is a loss function that measures the difference between the predicted value and the true output. The core idea behind classifier-based two-sample tests (Lopez-Paz & Oquab, 2017; Kim et al., 2021; Hediger et al., 2022) is that when the null hypothesis of equality of distributions is true, any classifier will return a random guess. On the other hand, when two distributions are significantly different, the accuracy of a reasonable classifier would be greater than chance level. Therefore, empirical classification accuracy can serve as an effective test statistic. However, since we do not observe a sample from f_{XY} but a sample from $f^{(2)}_{YX}$, we need to take this into consideration when we train a classifier. Specifically, we compute a classifier

$$\widehat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}\ell\big(h\big(V_i^{(1)}\big), 1\big) + \frac{1}{n_2}\sum_{i=1}^{n_2} r_X\big(X_i^{(2)}\big)\ell\big(h\big(V_i^{(2)}\big), 2\big)\right\}, \tag{4}$$

and use the empirical classification accuracy of $\widehat{h}$, again corrected by the density ratio, as our test statistic. When training and testing are performed on independent datasets, the asymptotic null distribution of the classification accuracy is approximately Gaussian (Kim et al., 2021; Hediger et al., 2022); thereby the critical value can be determined based on this Gaussian approximation. We provide a detailed analysis of this approach in Section 4.1, and present numerical results in Section 5.

Example 6. **(Kernel MMD)** The last example is a kernel MMD statistic (Gretton et al., 2012). Given a kernel k, the population MMD compares the kernel mean embeddings of two distributions with density functions $f_{YX}^{(1)}$ and f_{YX}, respectively. In a kernel form, the squared MMD can be written as

$$\mathrm{MMD}^2 = E\{k(V_1^{(1)}, V_2^{(1)})\} + E\{r_X(X_1^{(2)})r_X(X_2^{(2)})k(V_1^{(2)}, V_2^{(2)})\} - 2E\{r_X(X_2^{(2)})k(V_1^{(1)}, V_2^{(2)})\},$$

where the bias is corrected via importance weighting. The squared MMD can be estimated as

$$\frac{1}{n_1(n_1-1)} \sum_{1\le i\ne j\le n_1} k(V_i^{(1)}, V_j^{(1)}) + \frac{1}{n_2(n_2-1)} \sum_{1\le i\ne j\le n_2} r_X(X_i^{(2)})r_X(X_j^{(2)})k(V_i^{(2)}, V_j^{(2)}) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_X(X_j^{(2)})k(V_i^{(1)}, V_j^{(2)}),$$

which is an unbiased estimator of the population MMD. Unlike the other three test statistics mentioned earlier, this estimator converges to an infinite sum of weighted chi-squared distributions, whose weights are unknown in practice. This was not a major issue for marginal two-sample testing, where permutation tests could calibrate any test statistic in a nonparametric way. However, the standard permutation approach is no longer valid for conditional two-sample testing, which presents a challenge. Therefore, we focus on another estimator, a linear-time MMD statistic, in Section 4.2, which offers advantages in both the tractability of the asymptotic distribution and computational efficiency.

The preceding discussion assumes that the marginal density ratio r_X is known. As mentioned before, the success of this approach hinges on accurately estimating the density ratio r_X. To this end, one can draw upon a wide range of existing techniques in the literature for density ratio estimation (e.g., Sugiyama et al., 2010, 2012) to obtain a reliable testing result. Using the same dataset for both density ratio estimation and other parts of a statistic often results in plug-in bias. Hence, we recommend using auxiliary dataset obtained through, e.g., sample splitting, to estimate the density ratio. We concretely illustrate this approach in Section 4.1 and Section 4.2, using a classifier-based test statistic and a linear-time MMD statistic.

While various existing tools for density ratio estimation offer flexibility, an inherent drawback of this approach is that the behavior of the test statistic could be erratic when the density ratio is irregular and potentially unbounded. This issue can be mitigated by clipping the density ratio estimate at a certain value or shrinking it to zero (e.g., Shimodaira, 2000). Nevertheless, when prior knowledge indicates that the density ratio behaves poorly or is difficult to estimate, the testing performance may degrade (see Section 5 for numerical results) and thus this approach should be used with caution.

4.1. *Classifier-based Approach*

This subsection illustrates a classifier-based test for conditional two-sample testing. To simplify the presentation, we assume that $n_1 = n_2 = 2n$ and split the dataset into two: $D_a = \{V_i^{(1)}\}_{i=1}^n \cup \{V_i^{(2)}\}_{i=1}^n$ and $D_b = \{V_i^{(1)}\}_{i=n+1}^{2n} \cup \{V_i^{(2)}\}_{i=n+1}^{2n}$. For some positive integer $m < n$, we further divide D_a as $D_a^* = \{V_i^{(1)}\}_{i=1}^m \cup \{V_i^{(2)}\}_{i=1}^m$ and $D_a^{**} = D_a \setminus D_a^{(1)}$, and let $\widehat{r}_X$ denote an estimator of r_X formed on D_a^{**}. Additionally, let $\widehat{h}$ be a classifier trained as in (4) based on D_b. Let us

write $\widehat{A}_{1,i} = \mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\}$ and $\widehat{A}_{2,i} = \widehat{r}_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}$ for $i \in \{1, \ldots, m\}$, and define $\overline{A}_1 = m^{-1}\sum_{i=1}^m \widehat{A}_{1,i}$ and $\overline{A}_2 = m^{-1}\sum_{i=1}^m \widehat{A}_{2,i}$. The population-level classification accuracy of $\widehat{h}$ is $\mathrm{pr}\{\widehat{h}(V^{(1)}) = 1\}/2 + E[r_X(X^{(2)})\mathbb{1}\{\widehat{h}(V^{(2)}) = 2\}]/2$, which is $1/2$ for any classifier $\widehat{h}$ under the null hypothesis. This observation leads to a classifier-based test statistic for conditional two-sample testing given as

$$\widehat{\mathrm{Acc}} = \frac{m^{1/2}(\overline{A}_1 + \overline{A}_2 - 1)}{(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)^{1/2}}, \tag{5}$$

where $\widehat{\sigma}_1^2 = (m-1)^{-1}\sum_{i=1}^m (\widehat{A}_{1,i} - \overline{A}_1)^2$ and $\widehat{\sigma}_2^2 = (m-1)^{-1}\sum_{i=1}^m (\widehat{A}_{2,i} - \overline{A}_2)^2$. To formally establish the limiting distribution of $\widehat{\mathrm{Acc}}$, we consider the following assumptions.

Assumption 1. Let $m_n = m$ be an increasing sequence of positive integers with $\lim_{n\to\infty} m_n = \infty$. Consider a class of null distributions $\mathcal{P}_0$ such that

(a) There are constants $c_1, c_2 \in (0,1)$ such that $c_1 \le \inf_{P\in\mathcal{P}_0} \mathrm{pr}_P\{\widehat{h}(V^{(1)}) = 1 \mid \widehat{h}\} \le \sup_{P\in\mathcal{P}_0} \mathrm{pr}_P\{\widehat{h}(V^{(1)}) = 1 \mid \widehat{h}\} \le c_2$ for all sufficiently large n. Moreover, assume that there exist constants $C, \delta > 0$ such that $\sup_{P\in\mathcal{P}_0} E_P[\{\widehat{r}_X(X^{(2)})\}^{2+\delta}] \le C$ for all sufficiently large n.

(b) For any $\epsilon > 0$, the density ratio estimator satisfies

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \mathrm{pr}_P\big(mE_P[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2 \mid \widehat{r}_X] \ge \epsilon\big) = 0.$$

Assumption 1(a) is imposed to establish the (conditional) central limit theorem for the test statistic with the true density ratio, and it excludes a deterministic classifier, which would return the same prediction value regardless of inputs, under the null hypothesis. Assumption 1(b) ensures that the approximation error from $\widehat{r}_X$ is asymptotically negligible, which is similarly assumed in Hu & Lei (2024). In order to theoretically justify this condition, one needs to take m much smaller than the sample size used for training $\widehat{r}_X$. However, our empirical results in Assumption 5 illustrate that $\widehat{\mathrm{Acc}}$ approximates $N(0,1)$ closely even under balanced splitting. Therefore, echoing Hu & Lei (2024), we suggest taking $m = \lfloor n/2 \rfloor$ in practice.

Under Assumption 1, the classifier-based test statistic in (5) converges to $N(0,1)$ uniformly over $\mathcal{P}_0$.

Theorem 3. *For the class of null distributions $\mathcal{P}_0$ satisfying Assumption 1, $\widehat{\mathrm{Acc}}$ in (5) converges to $N(0,1)$:*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P(\widehat{\mathrm{Acc}} \le t) - \Phi(t)\right| = 0.$$

The proof of Theorem 3 can be found in supplementary material. According to Theorem 3, the classifier-based test rejects the null hypothesis when $\widehat{\mathrm{Acc}} > \Phi^{-1}(1-\alpha)$, which has asymptotic validity over $\mathcal{P}_0$ satisfying Assumption 1. We next improve the efficiency of this procedure via K-fold cross-validation. To describe the procedure, we begin by considering K disjoint subsets of D_a, denoted as $D_{a,1}, D_{a,2}, \ldots, D_{a,K}$, of equal size $m = \lfloor n/K \rfloor$ for simplicity. For $j \in \{1, \ldots, K\}$, let $\overline{A}_{1,j} + \overline{A}_{2,j} - 1$ and $\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2$ denote the quantities analogous to $\overline{A}_1 + \overline{A}_2 - 1$ and $\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$, respectively, by letting $D_a^* = D_{a,j}$ and $D_a^{**} = D_a \setminus D_{a,j}$. We then define the cross-validated

classification accuracy statistic as

$$\widehat{\mathrm{Acc}}_{\mathrm{cv}} = \frac{1}{K^{1/2}} \sum_{j=1}^{K} \frac{m^{1/2}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{(\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2)^{1/2}}. \tag{6}$$

The next corollary proves that the cross-validated accuracy statistic is asymptotically normally distributed under the null hypothesis.

Corollary 1. *Consider the same setting as in Theorem 3. For any fixed $K \geq 2$, it holds that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \left| \mathrm{pr}_P(\widehat{\mathrm{Acc}}_{\mathrm{cv}} \leq t) - \Phi(t) \right| = 0.$$

According to Corollary 1, the test that rejects the null when $\widehat{\mathrm{Acc}}_{\mathrm{cv}} > \Phi^{-1}(1-\alpha)$ maintains the asymptotic type I error under control. In terms of power, the cross-validated version is generally more powerful than the accuracy test without cross-validation as it uses the sample more efficiently. We numerically demonstrate this point in Section 5.

Many practical classifiers attempt to mimic the Bayes optimal classifier. For the balanced-sample setting, the Bayes optimal classifier is given as $h^\star(x,y) = \mathbb{1}\big(f_{YX}^{(1)}(y,x)/\{f_{YX}^{(1)}(y,x) + f_{YX}(y,x)\} > 1/2\big)$ whose classification accuracy can be explicitly computed in terms of the total variation (TV) distance. Specifically, twice the classification accuracy can be computed as

$$\mathrm{pr}\{h^\star(V^{(1)}) = 1\} + E\{r_X(X^{(2)})\mathbb{1}(h^\star(V^{(2)}) = 2)\} = 1 + \mathrm{TV}(f_{YX}^{(1)}, f_{YX}),$$

where $\mathrm{TV}(f_{YX}^{(1)}, f_{YX})$ denotes the TV distance between two distributions with densities $f_{YX}^{(1)}$ and f_{YX}, respectively. Since the TV distance becomes zero if and only if two distributions are identical, our classifier-based test can be powerful against general alternatives when the classifier in use approximates the Bayes classifier.

The next subsection develops parallel results using a linear-time MMD statistic.

4.2. *Linear-time MMD*

In this subsection, we provide a detailed treatment of our second framework by focusing on a linear-time MMD statistic (Gretton et al., 2012, Lemma 14) with a kernel k. As in Section 4.1, we assume that $n_1 = n_2 = 2n$ and split the dataset into two: $D_a = \{V_i^{(1)}\}_{i=1}^n \cup \{V_i^{(2)}\}_{i=1}^n$ and $D_b = \{V_i^{(1)}\}_{i=n+1}^{2n} \cup \{V_i^{(2)}\}_{i=n+1}^{2n}$. Letting $\widehat{r}_X$ be an estimator of r_X formed on D_b and $m = \lfloor n/2 \rfloor$, define

$$\begin{aligned}\widehat{S}_i = {} & k(V_i^{(1)}, V_{i+m}^{(1)}) + \widehat{r}_X(X_i^{(2)})\widehat{r}_X(X_{i+m}^{(2)})k(V_i^{(2)}, V_{i+m}^{(2)}) \\ & - \widehat{r}_X(X_i^{(2)})k(V_i^{(2)}, V_{i+m}^{(1)}) - \widehat{r}_X(X_{i+m}^{(2)})k(V_i^{(1)}, V_{i+m}^{(2)}).\end{aligned}$$

The test statistic that we analyze is a t-statistic based on $\widehat{S}_1, \ldots, \widehat{S}_m$. Specifically, letting $\overline{S} = m^{-1}\sum_{i=1}^m \widehat{S}_i$ and $\widehat{\sigma}^2 = (m-1)^{-1}\sum_{i=1}^m (\widehat{S}_i - \overline{S})^2$, the (studentized) linear-time MMD statistic is given as

$$\widehat{\mathrm{MMD}}_\ell^2 = \frac{m^{1/2}\overline{S}}{\widehat{\sigma}}. \tag{7}$$

In order to establish the asymptotic normality of $\widehat{\mathrm{MMD}}_\ell^2$, we make the following assumptions. Below, let S_i denote the quantity defined similarly as $\widehat{S}_i$ by replacing $\widehat{r}_X$ with the population counterpart r_X.

Assumption 2. Consider a class of null distributions $\mathcal{P}_0$ and assume that

(a) There exist constants $c, C > 0$ such that $\inf_{P\in\mathcal{P}_0} E_P(S_1^2) \geq c$ and $\sup_{P\in\mathcal{P}_0} E_P(S_1^{2+\delta}) \leq C$ for some $\delta > 0$.
(b) $\sup_{P\in\mathcal{P}_0} E_P\{(r_X(X^{(2)}))^2\} < \infty$ and $\sup_{P\in\mathcal{P}_0} E_P\left\{\left(\widehat{r}_X\left(X^{(2)}\right) - r_X\left(X^{(2)}\right)\right)^2\right\} = o(m^{-1/2})$.
(c) The kernel is uniformly bounded as $\|k\|_\infty \leq K$.

Assumption 2(a) is about the moment condition for the population counterpart of $\widehat{S}_i$. This assumption is required to apply the uniform central limit theorem. Assumption 2(b) is, on the other hand, required to prove that the difference between $\widehat{\mathrm{MMD}}_\ell^2$ using $\{\widehat{S}_i\}_{i=1}^m$ and that using $\{S_i\}_{i=1}^m$ are asymptotically negligible. This is similar to Assumption 1(b) and the assumption made in Hu & Lei (2024), but this condition is considerably weaker in terms of convergence rate. Assumption 2(c) assumes that the kernel k is uniformly bounded. While this assumption is met for many practical kernels (e.g., Gaussian kernel), it can be relaxed by adopting more complex moment or convergence assumptions.

Having stated the assumptions, we now present the asymptotic normality of $\widehat{\mathrm{MMD}}_\ell^2$ under the null hypothesis in (1).

Theorem 4. *For the class of null distributions $\mathcal{P}_0$ satisfying Assumption 2, $\widehat{\mathrm{MMD}}_\ell^2$ converges to $N(0,1)$ as*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P(\widehat{\mathrm{MMD}}_\ell^2 \leq t) - \Phi(t)\right| = 0.$$

We defer the proof of Theorem 4 to supplementary material. Based on Theorem 4, the test that rejects the null when $\widehat{\mathrm{MMD}}_\ell^2 > \Phi^{-1}(1-\alpha)$ controls the size uniformly over the class of distributions that satisfy Assumption 2.

Similarly to the classification-based test, we can improve the efficiency of $\widehat{\mathrm{MMD}}_\ell^2$ via K-fold cross-validation. To describe this, we begin by partitioning the dataset of size $2n$ into K-folds, denoted as $D_1, D_2, \ldots, D_K$, of equal size for simplicity. For $j \in \{1, \ldots, K\}$, let $\overline{S}_j$ and $\widehat{\sigma}_j^2$ denote the quantities similarly defined as $\overline{S}$ and $\widehat{\sigma}^2$, respectively, by letting $D_a = D_j$ and $D_b = \cup_{i=1}^K D_i \setminus D_j$. We then define the cross-validated MMD statistic as

$$\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2 = \frac{1}{K}\sum_{j=1}^K \frac{n^{1/2}\overline{S}_j}{\widehat{\sigma}_j}. \tag{8}$$

The next corollary shows that $\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2$ converges to $N(0,1)$ under the same conditions for Theorem 4.

Corollary 2. *Consider the same setting as in Theorem 4. Then for a fixed $K \geq 2$, it holds that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P(\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2 \leq t) - \Phi(t)\right| = 0.$$

Again, the test that rejects the null when $\widehat{\mathrm{MMD}}^2_{\mathrm{cv}} > \Phi^{-1}(1-\alpha)$ is asymptotically level α based on Corollary 2. When the kernel k is a characteristic kernel (e.g., Fukumizu et al., 2007), the population MMD becomes equal to zero if and only if two distributions coincide. Thus as for the classifier-based tests in Section 4.1, the MMD-based tests can be powerful against general alternatives, provided that the density ratio r_X can be accurately estimated.

The next section illustrates the numerical performance of the proposed tests in comparison to existing methods.

5. Numerical Experiments

In this section, we evaluate the empirical performance of the proposed tests, alongside existing methods from the literature, across various scenarios. Within each scenario, we compare the conditional independence testing (CIT) approach described in Section 3 with the density ratio-based testing (DRT) approach described in Section 4. Our experiments employ DRT methods including classifier-based tests in Section 4.1, linear-time MMD in Section 4.2, Conformal Prediction (CP) test (Hu & Lei, 2024), Debiased Conformal Prediction (DCP) test (Chen & Lei, 2024), Randomized Conditional Independence Test (RCIT) (Strobl et al., 2019), Generalized Covariance Measure (GCM) (Shah & Peters, 2020), Projected Covariance Measure (PCM) (Lundborg et al., 2022), and WGSC (Williamson et al., 2023). Detailed descriptions of these methods are provided in the supplementary material.

For DRT methods, we employ a classification-based approach using linear logistic regression for density ratio estimation (Sugiyama et al., 2010). For CIT methods relying on regression estimation, such as GCM and PCM, we use a Random Forest as the underlying regression model. For the WGSC, we use XGBoost, as it performs best in our experiments. All the results are averaged over 500 repetitions with a significance level $\alpha = 0.05$.

It is crucial to note that the efficacy and validity of these methods depend on specific assumptions, which vary across different approaches. For example, DRT methods rely heavily on accurate density ratio estimation, while CIT methods depend on reliable estimation of conditional operators, such as regression functions. We aim to emphasize this distinction by providing a comprehensive evaluation of these approaches for conditional two-sample testing. This analysis sheds light on how each method performs under different conditions, guiding practitioners in choosing the most suitable approach for their specific applications.

5.1. *Synthetic Data Examples*

We design three synthetic data scenarios to evaluate the performance of conditional two-sample testing methods under different conditions. Each scenario is implemented with both unbounded (U) and bounded (B) density ratios to assess how the difficulty of density ratio estimation affects the performance of each method. The marginal distributions of X remain consistent across scenarios with unbounded density ratios, and similarly, across scenarios with bounded density ratios. In the examples, the dimension of the covariates X is fixed at $p = 10$.

For the unbounded case (U), we employ Gaussian distributions for marginal distributions of X. Specifically, for $k = 1$, samples are drawn from a standard Gaussian distribution, i.e., $x^{(1)} \sim N(0, I_p)$, where I_p is the identity matrix of dimension p. For $k = 2$, we introduce a covariate shift by sampling from a Gaussian distribution with mean vector $\mu = (1, 1, -1, -1, 0, \ldots, 0)^\top$

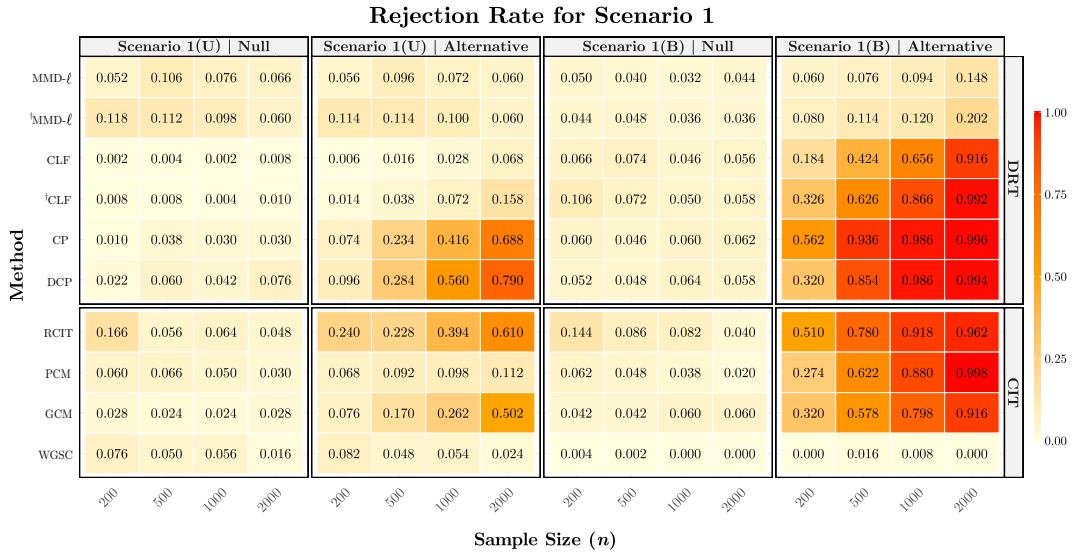

Fig. 1: Rejection rates for Scenario 1 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

and the same covariance structure, i.e., $x^{(2)} \sim N(\mu, I_p)$. In the bounded case (B), we truncate the support of both distributions to $[-0.5, 0.5]$ in each dimension, resulting in truncated Gaussian distributions, $x^{(1)} \sim TN(0, I_p)$ and $x^{(2)} \sim TN(\mu, I_p)$, where μ is the same as in the unbounded case.

Scenario 1: Linear Model with Mean Shift. Inspired by Hu & Lei (2024), this scenario investigates the efficacy of testing methods in detecting the mean difference between two linear models. For each $k \in \{1, 2\}$, we set $y^{(j)} \mid x^{(j)} = \delta^{(j)} + x^{(j)\top}\beta + \epsilon^{(j)}$, where $\epsilon^{(j)}$ follows a t-distribution with 2 degrees of freedom. The regression coefficient β is set to $(1, -1, -1, 1, 0, \ldots, 0)^\top$. Under the null hypothesis, we set $\delta^{(1)} = \delta^{(2)} = 0$, while for the alternative hypothesis, we set $\delta^{(1)} = 0$ and $\delta^{(2)} = 0.5$.

Scenario 2: High Variability in Conditional Distribution. We investigate the effect of high variability in the conditional distribution, slightly modifying the example from Chatterjee et al. (2024, Section 6.2). Under the null hypothesis, we model the conditional distributions as $y^{(j)} \mid x^{(j)} \sim N\left(x^{(j)\top}\beta^{(j)}, (\sigma^{(j)})^2\right)$, where $\beta^{(j)} = \mathbf{1}_p$ defined as a p-dimensional vector of ones and $(\sigma^{(j)})^2 = 10^2$ for both $j \in \{1, 2\}$. For the alternative hypothesis, we modify $\beta^{(2)}$ to $(1, \ldots, 1, 0)^\top$ and introduce heteroscedasticity by varying the variance for $j = 2$ as $(\sigma^{(2)})^2 = 10\left(1 + \exp\left(-\|x^{(2)} - 0.5\mathbf{1}_p\|_2^2/64\right)\right)$.

Scenario 3: Post-Nonlinear Model. Our final scenario considers a post-nonlinear (PNL) model, widely used in causal predictive inference (Zhang et al., 2017; Li et al., 2023). We model the conditional distributions as $y^{(j)} \mid x^{(j)} = f^{(j)}(x^{(j)\top}\mathbf{1}_p + 2\epsilon)$, where $\epsilon \sim N(0, 1)$ and $k \in \{1, 2\}$. Under the null hypothesis, we set $f^{(j)}(x) = \cos(x)$ for both $j \in \{1, 2\}$, while for the alternative hypothesis, $f^{(2)}(x)$ is randomly sampled from the set $\{x, x^2, x^3, \sin(x), \tanh(x)\}$.

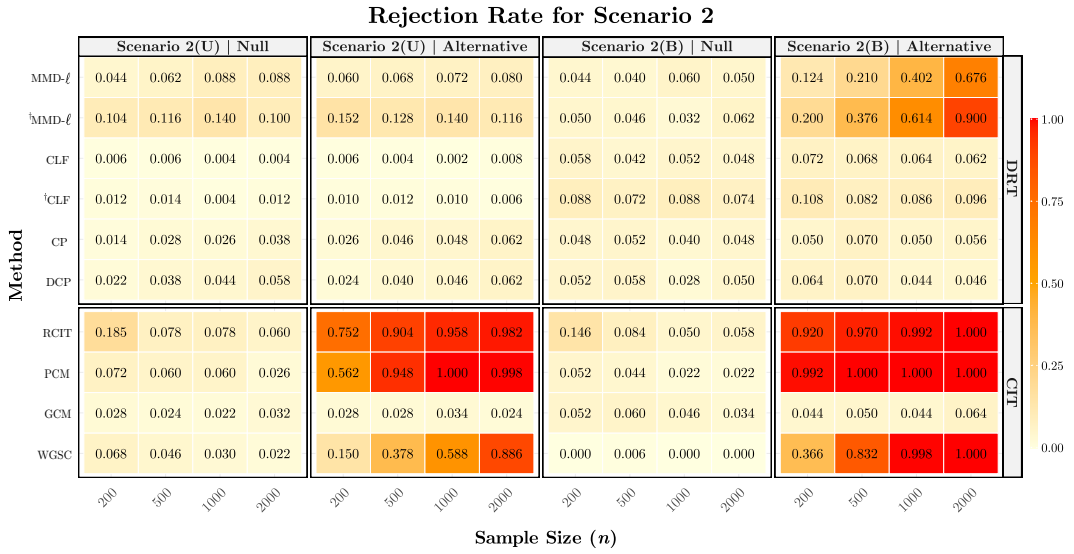

Fig. 2: Rejection rates for Scenario 2 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

Fig. 3: Rejection rates for Scenario 3 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

Our experimental results provide several key insights into the performance of conditional two-sample testing methods across diverse scenarios. A consistent pattern observed throughout all scenarios is the superior performance of DRT methods in bounded settings compared to unbounded settings. This improvement can be attributed to the relative ease of density ratio estimation when the density ratio is bounded, leading to more stable results. In contrast, CIT

methods exhibit relatively consistent performance regardless of whether the density ratios are bounded or unbounded.

In Scenario 1, the classifier-based test shows the most significant improvement in performance when transitioning from unbounded to bounded cases. However, MMD-ℓ shows lower sensitivity in detecting mean shifts compared to other DRT methods. Among CIT methods, RCIT and GCM exhibit the best performance in this scenario. Scenario 2 highlights the strengths of MMD-ℓ, which only considers the marginal density ratio of X, in comparison to other DRT methods that account for conditional density ratios. MMD-ℓ shows a distinct advantage in this scenario. CIT methods also generally perform well under these conditions, showing their robustness to complex distributional changes. In Scenario 3, which tests the ability of methods to detect non-linear relationships, all DRT methods improve performance in the bounded case. Among CIT methods, there is no significant difference in performance, except for GCM and WGSC. GCM shows improved performance in the bounded case, while WGSC shows degraded performance.

These results underscore the critical role of accurate density ratio estimation in determining the performance of DRT methods. While CIT methods demonstrate consistent performance across both bounded and unbounded cases, suggesting their utility in a wide range of practical scenarios, they also have limitations. CIT methods, particularly regression-based approaches like GCM, PCM, and WGSC, can be sensitive to the choice of regression model, as we demonstrate in the supplementary material. Notably, RCIT exhibits high type I error rates in all scenarios when sample sizes are relatively small, suggesting that caution is needed when applying RCIT to limited datasets. On the other hand, some methods show overly conservative behavior in certain scenarios. The cross-validated versions of DRT methods (†MMD-ℓ and †CLF) consistently show power gains compared to their non-cross-validated counterparts as discussed in Section 4.1 and Section 4.2. Overall, our findings offer important insights into the strengths and limitations of different conditional two-sample testing methods.

5.2. *Real Data Analysis*

We further evaluate the performance of our proposed approaches on two real-world datasets: the diamonds dataset and the superconductivity dataset. Following Kim et al. (2023), we treat each dataset as a population from which we draw samples, allowing for controlled experiments with known ground truth. Prior to analysis, we apply standard scaling to both X and Y variables. To introduce covariate shift, we implement biased sampling procedures. Specifically, we sample $X^{(1)}$ uniformly from the original feature space, while $X^{(2)}$ is sampled with probability proportional to $\exp(-x_1^2)$, where x_1 denotes the first feature of X. For the response variable Y, under the null hypothesis, we employ uniform sampling for both $Y^{(1)}$ and $Y^{(2)}$. Under the alternative hypothesis, $Y^{(1)}$ is sampled uniformly, while $Y^{(2)}$ is sampled with probability proportional to $\exp(-y)$, where y represents the values of Y in the dataset. Fig. 4 illustrates the performance of the DRT methods on both datasets, using linear logistic (LL) and kernel logistic regression (KLR) for density ratio estimation.

Diamonds dataset. The diamonds dataset, available in the R package `ggplot2`, consists of 53,490 observations and 10 features, including price, carat, clarity and color. In our analysis, we set the price variable as Y, and use the 6 numerical variables `(carat, depth, table, x, y, z)` as X. As illustrated in Fig. 4, most DRT methods exhibit good type I error control

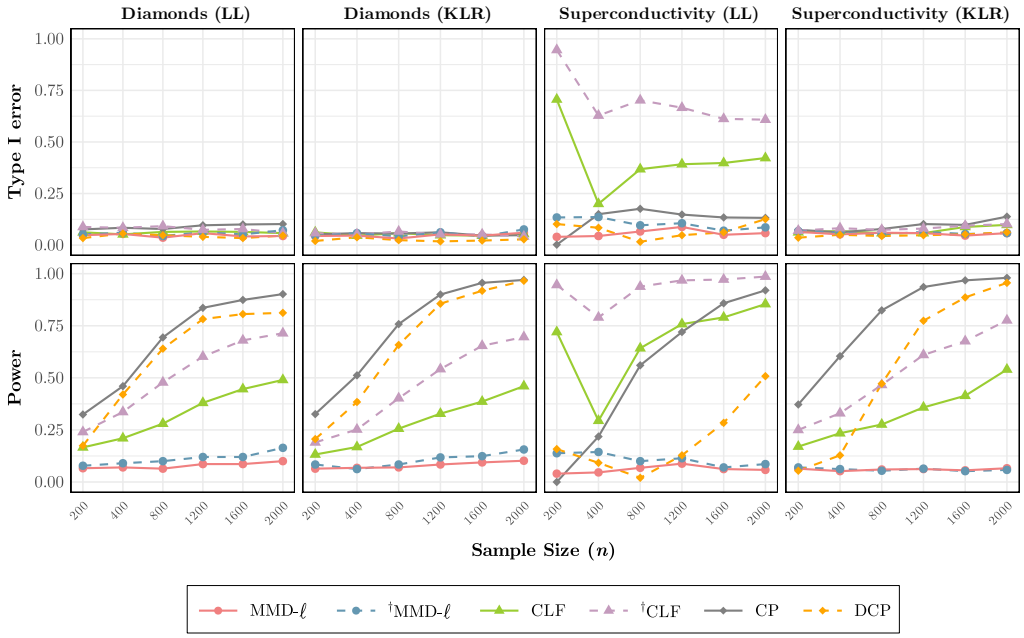

Fig. 4: Performance comparison of DRT methods on diamonds and superconductivity datasets using LL and KLR for density ratio estimation. Rejection rates are averaged over 500 repetitions with $\alpha = 0.05$, under null (*top*) and alternative (*bottom*) hypotheses.

under both LL and KLR, with rejection rates generally close to the significance level α. Under the alternative hypothesis, we observe a clear trend of increasing power with sample size for all methods. Particularly, the cross-validated versions (†MMD-ℓ and †CLF) exhibit improved power, consistent with our observations in the synthetic data examples.

Superconductivity dataset. The superconductivity dataset, obtained from the UCI Machine Learning Repository and compiled by Hamidieh (2018), presents a more complex and high-dimensional challenge compared to the diamonds dataset. It comprises 81 features extracted from 21,263 superconductors, with the critical temperature at which the material transitions to a superconducting state serving as the response variable Y. The results reveal a significant contrast between density ratio estimation methods based on LL and KLR. Under LL, several DRT methods, especially the classifier-based tests, struggle to control the type I error, with rejection rates far exceeding the significance level. Conversely, when using KLR for density ratio estimation, DRT methods show improved type I error control.

These empirical findings emphasize the importance of carefully considering the nature of the data and the choice of density ratio estimation techniques when applying DRT methods for conditional two-sample testing. The performance of different methods can vary significantly, indicating the need for careful method selection and, potentially, more advanced approaches when handling complex and high-dimensional data. While we focus on DRT methods in this section, experimental results for CIT methods are presented in the supplementary material for completeness.

6. Discussion

In this paper, we shed new light on the relatively underexplored problem of conditional two-sample testing. We begin by characterizing the fundamental difficulty of the problem and highlighting the importance of assumptions to make it feasible. We then introduce two general frameworks: (1) converting conditional independence tests into conditional two-sample tests and (2) transforming the problem of comparing conditional distributions into marginal distributions based on density ratio estimation. Both approaches offer significant flexibility, allowing one to leverage well-developed tools to effectively tackle the problem.

Our work opens up several interesting directions for future work. One promising avenue is to extend our framework to conditional K-sample testing with a general $K \geq 2$. Such an extension would expand the applicability of our framework beyond the comparison of just two groups. This setting is related to conditional independence testing where Z is a categorical random variable taking values in $\{1, 2, \ldots, K\}$. We expect our results established in Section 3 to serve as a cornerstone for this extension. Another direction worth exploring is establishing a framework for conditional two-sample testing based on resampling methods. One promising approach is the Sampling Importance Resampling (SIR) algorithm (Givens & Hoeting, 2012, Chapter 6.3), which allows us to obtain an approximate sample from the distribution with density f_{YX}. Future work can focus on methods that compare the sample from $P_{XY}^{(1)}$ with the approximate sample obtained from the SIR algorithm. Finally, one can explore other two-sample test statistics beyond those listed in Section 4. Of particular interest is the block-wise MMD statistic (Zaremba et al., 2013). This statistic has a tractable limiting distribution, while achieving lower variance with a slight increase in computational cost compared to the linear-time MMD statistic. We leave all these interesting topics for future work.

Acknowledgement

We acknowledge support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), and the Korea government (MSIT) RS-2023-00211073. We are grateful to Gyumin Lee for the careful proofreading.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of the main results, supporting lemmas and details of numerical experiments and additional simulation results.

References

Andrews, D. W. K. (1997). A Conditional Kolmogorov Test. *Econometrica* **65**, 1097–1128.

Barocas, S., Hardt, M. & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press.

Berrett, T. B., Wang, Y., Barber, R. F. & Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 175–197.

Boeken, P. A. & Mooij, J. M. (2021). A bayesian nonparametric conditional two-sample test with an application to local causal discovery. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, C. de Campos & M. H. Maathuis, eds., vol. 161 of *Proceedings of Machine Learning Research*. PMLR.

Candes, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**, 551–577.

Chakraborty, A., Zhang, J. & Katsevich, E. (2024). Doubly robust and computationally efficient high-dimensional variable selection. *arXiv preprint arXiv:2409.09512* .

Chatterjee, A., Niu, Z. & Bhattacharya, B. B. (2024). A kernel-based conditional two-sample test using nearest neighbors (with applications to calibration, regression curves, and simulation-based inference). *arXiv preprint arXiv:2407.16550* .

Chen, Y. & Lei, J. (2024). De-Biased Two-Sample U-Statistics With Application To Conditional Distribution Testing. *arXiv preprint arXiv:2402.00164* .

Choi, K., Liao, M. & Ermon, S. (2021). Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*.

Choi, K., Meng, C., Song, Y. & Ermon, S. (2022). Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*.

Chung, E. & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* **41**, 484–507.

Dai, B., Shen, X. & Pan, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE transactions on neural networks and learning systems* **35**, 1898–1911.

Doran, G., Muandet, K., Zhang, K. & Schölkopf, B. (2014). A permutation-based kernel conditional independence test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14. Arlington, Virginia, USA: AUAI Press.

Fan, J. & Lin, S.-K. (1998). Test of significance when data are curves. *Journal of the American Statistical Association* **93**, 1007–1021.

Fan, Y., Li, Q. & Min, I. (2006). A nonparametric bootstrap test of conditional distributions. *Econometric Theory* **22**, 587–613.

Fukumizu, K., Gretton, A., Sun, X. & Schölkopf, B. (2007). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems* **20**, 489–496.

Givens, G. H. & Hoeting, J. A. (2012). *Computational statistics*. Hoboken, NJ, USA: John Wiley & Sons, 2nd ed.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13**, 723–773.

Hall, P. & Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* **85**, 1039–1049.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science* **154**, 346–354.

Hardt, M., Price, E., Price, E. & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon & R. Garnett, eds., vol. 29. Curran Associates, Inc.

Hediger, S., Michel, L. & Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis* **170**, 107435.

Hu, X. & Lei, J. (2024). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association* **119**, 1136–1154.

Kanamori, T., Hido, S. & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research* **10**, 1391–1445.

Kanamori, T., Suzuki, T. & Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences* **93**, 787–798.

Kim, I., Balakrishnan, S. & Wasserman, L. (2022a). Minimax optimality of permutation tests. *The Annals of Statistics* **50**, 225–251.

Kim, I., Lee, A. B. & Lei, J. (2019). Global and local two-sample tests via regression. *Electronic Journal of Statistics* **13**, 5253–5305.

Kim, I., Neykov, M., Balakrishnan, S. & Wasserman, L. (2022b). Local permutation tests for conditional independence. *The Annals of Statistics* **50**, 3388–3414.

Kim, I., Neykov, M., Balakrishnan, S. & Wasserman, L. (2023). Conditional Independence Testing for Discrete Distributions: Beyond χ^2-and G-tests. *arXiv preprint arXiv:2308.05373* .

Kim, I., Ramdas, A., Singh, A. & Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics* **49**, 411–434.

Kimura, M. & Hino, H. (2024). A short survey on importance weighting for machine learning. *arXiv preprint arXiv:2403.10175* .

Kulasekera, K. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association* **90**, 1085–1093.

Kulasekera, K. & Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association* **92**, 500–511.

Li, S., Zhang, Y., Zhu, H., Wang, C., Shu, H., Chen, Z., Sun, Z. & Yang, Y. (2023). K-nearest-neighbor local sampling based conditional independence testing. *Advances in Neural Information Processing Systems* **36**, 23321–23344.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A. & Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*.

Liu, M., Katsevich, E., Janson, L. & Ramdas, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109**, 277–293.

Liu, S., Takeda, A., Suzuki, T. & Fukumizu, K. (2017). Trimmed density ratio estimation. *Advances in Neural Information Processing Systems* **30**, 4521–4531.

Lopez-Paz, D. & Oquab, M. (2017). Revisiting Classifier Two-Sample Tests. In *International Conference on Learning Representations*.

Lundborg, A. R., Kim, I., Shah, R. D. & Samworth, R. J. (2022). The Projected Covariance Measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039 (accepted to the Annals of Statistics)* .

Neumeyer, N. & Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics* **31**, 880–920.

Neykov, M., Balakrishnan, S. & Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics* **49**, 2151–2177.

Neykov, M., Wasserman, L., Kim, I. & Balakrishnan, S. (2023). Nearly Minimax Optimal Wasserstein Conditional Independence Testing. *arXiv preprint arXiv:2308.08672* .

Pardo-Fernández, J. C., Jiménez-Gamero, M. D. & El Ghouch, A. (2015). Tests for the equality of conditional variance functions in nonparametric regression. *Electronic Journal of Statistics* **9**.

Pogodin, R., Schrab, A., Li, Y., Sutherland, D. J. & Gretton, A. (2024). Practical Kernel Tests of Conditional Independence. *arXiv preprint arXiv:2402.13196* .

Rhodes, B., Xu, K. & Gutmann, M. U. (2020). Telescoping density-ratio estimation. *Advances in Neural Information Processing Systems* **33**, 4905–4916.

Scheidegger, C., Hörrmann, J. & Bühlmann, P. (2022). The weighted generalised covariance measure. *Journal of Machine Learning Research* **23**, 1–68.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B. & Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research* **24**, 1–81.

Shah, R. D. & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* **48**, 1514–1538.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90**, 227–244.

Strobl, E. V., Zhang, K. & Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference* **7**, 20180017.

Sugiyama, M., Krauledat, M. & Müller, K.-R. (2007a). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **8**, 985–1005.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. & Kawanabe, M. (2007b). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems* **20**.

Sugiyama, M., Suzuki, T. & Kanamori, T. (2010). Density ratio estimation: A comprehensive review. *RIMS Kokyuroku* , 10–31.

Sugiyama, M., Suzuki, T. & Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R. & Blei, D. M. (2022). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics* **31**, 151–162.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S. & Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing* **17**, 138–155.

Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. (2022). Gene–environment interactions and their impact on human health. *Genes & Immunity* **24**, 1–11.

Williamson, B. D., Gilbert, P. B., Simon, N. R. & Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association* **118**, 1645–1658.

Wu, H., Eckhardt, C. M. & Baccarelli, A. A. (2023). Molecular mechanisms of environmental exposures and human disease. *Nature Reviews Genetics* **24**, 332–344.

Yan, J. & Zhang, X. (2022). A nonparametric two-sample conditional distribution test. *arXiv preprint arXiv:2210.08149* .

Zaremba, W., Gretton, A. & Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems* **26**.

Zhang, K., Huang, B., Zhang, J., Glymour, C. & Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: skeleton estimation and orientation determination. *Proceedings of the 26th International Joint Conference on Artificial Intelligence* , 1347–1353.

Zhang, K., Peters, J., Janzing, D. & Schölkopf, B. (2011). Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11. Arlington, Virginia, USA: AUAI Press.

Zheng, J. X. (2000). A consistent test of conditional parametric distributions. *Econometric Theory* **16**, 667–691.