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

Supplementary material for "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

Summary

This supplementary file contains proofs of the theoretical results and additional simulations.

Overview of supplementary material. In Section 1, we present the proofs that include all the results discussed in the main paper. Section 2 gathers several lemmas that support these proofs. Finally, Section 3 provides implementation details of numerical experiments and additional simulation results.

1. Proofs

Notation. For real sequences (a_n) and (b_n), we say that $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all n. Let $(X_{P,n})_{n\in\mathbb{N}, P\in\mathcal{P}}$ be a family of sequences of random variables determined by $P \in \mathcal{P}$. We say that $X_{P,n} = X_n = o_{\mathcal{P}}(n^{-a})$ if for all $\epsilon > 0$,

$$\sup_{P\in\mathcal{P}} \mathrm{pr}_P(n^a |X_{P,n}| > \epsilon) \to 0.$$

For a positive integer n, we use the shorthand $[n]$ to denote the set $\{1, \ldots, n\}$.

1.1. *Proof of Theorem 1*

Proof of Theorem 1. For a constant $N \in \mathbb{N}$ greater than $2n$, we work with N i.i.d. random samples $\{(X_i, Y_i, Z_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ and define $N_1' = \sum_{i=1}^N \mathbb{1}(Z_i = 1)$ and $N_2' = \sum_{i=1}^N \mathbb{1}(Z_i = 2)$, which follow $N_1' \sim \text{Binomial}(N, \lambda_n)$ and $N_2' \sim \text{Binomial}(N, 1 - \lambda_n)$, respectively. For $n_1, n_2 \in \mathbb{N}$ given in the

theorem statement, define a good event $\mathcal{A} = \{N_1' \geq n_1, N_2' \geq n_2\}$, whose probability satisfies $\mathrm{pr}(\mathcal{A}) \geq 1 - \mathrm{pr}(N_1' < n_1) - \mathrm{pr}(N_2' < n_2)$ by the union bound. Since $N \geq 2n$, we can ensure that $n_1 - N\lambda_n \leq -\frac{1}{2}N\lambda_n$ and thus

$$\begin{aligned}\mathrm{pr}(N_1' < n_1) &= \mathrm{pr}(N_1' - N\lambda_n < n_1 - N\lambda_n) \leq \mathrm{pr}\Big(N_1' - N\lambda_n < -\frac{1}{2}N\lambda_n\Big) \\ &\leq \mathrm{pr}\Big(\big|N_1' - N\lambda_n\big| > \frac{1}{2}N\lambda_n\Big) \leq \frac{4(1-\lambda_n)}{N\lambda_n},\end{aligned}$$

where the last inequality uses Chebyshev's inequality along with $N_1' \sim \mathrm{Binomial}(N, \lambda_n)$. We can similarly obtain that $\mathrm{pr}(N_2' < n_2) \leq \frac{4\lambda_n}{N(1-\lambda_n)}$. Therefore, the probability of the good event $\mathcal{A}$ is lower bounded as

$$\mathrm{pr}(\mathcal{A}) \geq 1 - \frac{4(1-\lambda_n)^2 + 4\lambda_n^2}{N\lambda_n(1-\lambda_n)} \overset{\text{set}}{=} 1 - \varepsilon_N, \tag{1}$$

where $\varepsilon_N \to 0$ as $N \to \infty$.

Now consider a test ϕ that only uses $n_1 + n_2$ data points out of N samples. Importantly, this sample consists of n_1 observations from $\{(Y_i, X_i) : Z_i = 1\}$ and n_2 observations from $\{(Y_i, X_i) : Z_i = 2\}$, whenever $\mathbb{1}(\mathcal{A}) = 1$. If $\mathbb{1}(\mathcal{A}) = 0$, this test simply returns 0 (i.e., accept H_0). Moreover, we assume that this test satisfies $\sup_{P \in \mathcal{P}_{0,M}} E_P(\phi \mid \mathcal{A}) \leq \alpha$. In fact, since ϕ only uses n_1 observations with $Z = 1$ and n_2 observations with $Z = 2$, the previous inequality implies $\sup_{P \in \mathcal{P}_{0,M}} E_P(\phi \mid N_1 = n_1, N_2 = n_2) \leq \alpha$, i.e., ϕ is a valid level α test for conditional two-sample testing. Based on the previous results, the type I error of ϕ constructed based on $\{(X_i, Y_i, Z_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ fulfills

$$\sup_{P \in \mathcal{P}_{0,M}} E_P(\phi) = \sup_{P \in \mathcal{P}_{0,M}} E_P\{\phi \mathbb{1}(\mathcal{A})\} \leq \sup_{P \in \mathcal{P}_{0,M}} E_P(\phi \mid \mathcal{A}) \leq \alpha.$$

This implies that ϕ is a valid test for conditional independence with size α. Therefore, for any $P \in \mathcal{P}_{1,M}$,

$$\begin{aligned}&E_P(\phi \mid \mathcal{A})(1 - \varepsilon_N) \overset{\text{(i)}}{\leq} E_P(\phi \mid \mathcal{A}) E_P\{\mathbb{1}(\mathcal{A})\} \leq E_P(\phi) \overset{\text{(ii)}}{\leq} \alpha \\ \iff\ &E_P(\phi \mid \mathcal{A}) \overset{\text{(iii)}}{=} E_P(\phi \mid N_1 = n_1, N_2 = n_2) \leq \frac{\alpha}{1 - \varepsilon_N},\end{aligned}$$

where step (i) uses the inequality in (1), step (ii) holds by Shah & Peters (2020, Theorem 2 and Remark 4) and step (iii) uses the fact that ϕ only uses $n_1 + n_2$ observations as described before. Since N was an arbitrary number greater than or equal to $2n$ and $\varepsilon_N \to 0$ as $N \to \infty$, we can conclude that $E_P(\phi \mid N_1 = n_1, N_2 = n_2) \leq \alpha$ for any $P \in \mathcal{P}_{1,M}$.

1.2. *Proof of Theorem 2*

Proof of Theorem 2. We may write $\tilde{\phi} = \mathbb{1}(\tilde{n}_1 \leq n_1)\mathbb{1}(\tilde{n}_2 \leq n_2)\phi_{\tilde{n}}$ where $\phi_{\tilde{n}}$ is defined as ϕ based on $\mathcal{D}_{\tilde{n}} = \{(X_i, Y_i, Z_i)\}_{i=1}^{\tilde{n}}$ in Algorithm 1. Now generate new i.i.d. samples $\tilde{\mathcal{D}}_{\tilde{n}} = \{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^{\tilde{n}} \overset{\text{i.i.d.}}{\sim} P$ independent of $\mathcal{D}_{\tilde{n}}$ and define as

$$\tilde{\phi}^\dagger = \mathbb{1}\Big\{\sum_{i=1}^{\tilde{n}} \mathbb{1}(\tilde{Z}_i = 1) \leq n_1\Big\}\mathbb{1}\Big\{\sum_{i=1}^{\tilde{n}} \mathbb{1}(\tilde{Z}_i = 2) \leq n_2\Big\}\phi_{\tilde{n}}^\dagger,$$

where $\phi^{\dagger}_{\tilde{n}}$ is defined as ϕ but based on $\tilde{\mathcal{D}}_{\tilde{n}}$. Observe that the conditional distribution of $\tilde{\phi}$ given $N_1 = n_1, N_2 = n_2$ is the identical to the marginal distribution of $\tilde{\phi}^{\dagger}$. Thus it can be seen that

$$E_P\big(\tilde{\phi} \mid N_1 = n_1, N_2 = n_2\big) = E_P(\tilde{\phi}^{\dagger}) \leq E_P(\phi^{\dagger}_{\tilde{n}}),$$

for all P. In other words, we can effectively remove conditioning on $N_1 = n_1, N_2 = n_2$ by working with $\tilde{\phi}^{\dagger}$. Therefore, the first claim on type I error control follows.

Moving to the type II error, observe that

$$\begin{aligned}\mathbb{1}(\tilde{n}_1 \leq n_1)\mathbb{1}(\tilde{n}_2 \leq n_2)\phi_{\tilde{n}} &= \phi_{\tilde{n}} - \mathbb{1}(\tilde{n}_1 > n_1 \text{ or } \tilde{n}_2 > n_2)\phi_{\tilde{n}} \\ &\geq \phi_{\tilde{n}} - \mathbb{1}(\tilde{n}_1 > n_1) - \mathbb{1}(\tilde{n}_2 > n_2),\end{aligned}$$

by the union bound, which yields

$$E_P(1 - \tilde{\phi} \mid N_1 = n_1, N_2 = n_2) = E_P(1 - \tilde{\phi}^{\dagger}) \leq E_P(1 - \phi^{\dagger}_{\tilde{n}}) + E_P\{\mathbb{1}(\tilde{n}_1 > n_1)\} + E_P\{\mathbb{1}(\tilde{n}_2 > n_2)\}. \tag{2}$$

In addition, letting $p = n_1/n$ in Lemma 1 of Section 2, take $(1+\delta)\tilde{n}n_1/n = k(1+\delta)n_1 = n_1$, which gives $1 + \delta = k^{-1} \iff \delta = k^{-1} - 1$. Thus, by Lemma 1, we have

$$E_P\{\mathbb{1}(\tilde{n}_1 > n_1)\} \leq \exp\left\{-\frac{n_1 k(k^{-1}-1)^2}{3}\right\}.$$

Letting the right-hand side equal ε and solving for $k \in (0,1)$, we derive the form of k^* as presented in Algorithm 1, which shows that $E_P\{\mathbb{1}(\tilde{n}_1 > n_1)\} \leq \varepsilon$. By symmetry, the same analysis holds for the inequality $E_P\{\mathbb{1}(\tilde{n}_2 > n_2)\} \leq \varepsilon$. As a result, continuing from the inequality (2), we can upper bound the type II error of $\tilde{\phi}_{\alpha}$ as

$$E_P(1 - \tilde{\phi} \mid N_1 = n_1, N_2 = n_2) \leq E_P(1 - \phi^{\dagger}_{\tilde{n}}) + 2\varepsilon.$$

Since $\varepsilon = o(1)$, the above display proves the second claim on type II error control. This completes the proof of Theorem 2.

1.3. *Proof of Theorem 3*

We analyze the numerator and the denominator of $\widehat{\text{Acc}}$, separately. In particular, we first show that the numerator converges to a Gaussian distribution and the denominator is ratio-consistent to the population-level standard deviation under Assumption 1.

Analysis of the numerator. Starting with the numerator, let us rewrite

$$\begin{aligned}\overline{A}_1 + \overline{A}_2 - 1 &= \frac{1}{m}\sum_{i=1}^{m}\Big[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1\Big] \\ &\qquad + \frac{1}{m}\sum_{i=1}^{m}\big\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\big\}\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} \\ &= \frac{1}{m}\sum_{i=1}^{m}\underbrace{\Big[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1\Big]}_{=L_i(\widehat{h})} + o_{\mathcal{P}_0}(m^{-1/2}),\end{aligned}$$

where the last approximation holds since

$$\left|\frac{1}{m}\sum_{i=1}^{m}\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}\right| \le \left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}^2\right]^{1/2},$$

and the upper bound is $o_{\mathcal{P}_0}(m^{-1/2})$ due to Assumption 1(b). Thus $\overline{A}_1 + \overline{A}_2 - 1$ is dominated by the average of $L_i(\widehat{h})$ values. Given this and Slutsky's theorem, it suffices to study the limiting distribution of the sample average of $L_i(\widehat{h})$. Indeed, under Assumption 1(a), the conditional central limit theorem (Lemma 3) yields that

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}}\left|\mathrm{pr}_P\left(\frac{\frac{1}{m^{1/2}}\sum_{i=1}^{m}L_i(\widehat{h})}{\{\mathrm{var}[L(\widehat{h}) \mid \widehat{h}]\}^{1/2}} \le t\right) - \Phi(t)\right| = 0.$$

Consistency of the variance estimate. We next show the ratio-consistency of the variance estimator. Observe that

$$\mathrm{var}[L(\widehat{h}) \mid \widehat{h}] = \underbrace{\mathrm{var}[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} \mid \widehat{h}]}_{=\sigma_1^2} + \underbrace{\mathrm{var}\big[r_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}\big]}_{=\sigma_2^2},$$

and

$$\left|\frac{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}{\sigma_1^2 + \sigma_2^2} - 1\right| \le \left|\frac{\widehat{\sigma}_1^2 - \sigma_1^2}{\sigma_1^2}\right| + \left|\frac{\widehat{\sigma}_2^2 - \sigma_2^2}{\sigma_2^2}\right|.$$

Therefore, in order to show the ratio consistency of $\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$, it suffices to show the ratio consistency of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, individually. To this end, we use conditional Chebyshev's inequality and show

$$\sup_{P\in\mathcal{P}_0}\mathrm{pr}_P\big(|\widehat{\sigma}_1^2/\sigma_1^2 - 1| \ge t \mid \widehat{h}\big) \le \frac{1}{t^2}\sup_{P\in\mathcal{P}_0}\mathrm{var}_P(\widehat{\sigma}_1^2/\sigma_1^2 \mid \widehat{h}) \le \frac{1}{t^2 c_1(1-c_2)m}.$$

for sufficiently large n and for all $t > 0$, under Assumption 1(a). Hence $\widehat{\sigma}_1^2/\sigma_1^2$ converges to one in probability uniformly over $\mathcal{P}_0$. On the other hand, letting $A_{2,i} = \widehat{r}_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}$, we have

$$\begin{aligned}
\widehat{\sigma}_2^2 &= \frac{1}{m-1}\sum_{i=1}^{m}\left(A_{2,i} - \frac{1}{m}\sum_{j=1}^{m}A_{2,j}\right)^2 \\
&\quad + \frac{1}{m-1}\sum_{i=1}^{m}\left\{(\widehat{A}_{2,i} - A_{2,i}) - \frac{1}{m}\sum_{j=1}^{m}(\widehat{A}_{2,j} - A_{2,j})\right\}^2 \\
&\quad + \frac{2}{m-1}\sum_{i=1}^{m}\left(A_{2,i} - \frac{1}{m}\sum_{j=1}^{m}A_{2,j}\right)\left[(\widehat{A}_{2,i} - A_{2,i}) - \frac{1}{m}\sum_{j=1}^{m}\{\widehat{A}_{2,j} - A_{2,j}\}\right] \\
&= (\mathbb{I}) + (\mathbb{III}) + (\mathbb{IIII}).
\end{aligned}$$

Similarly as before, the term $(\mathbb{I})/\sigma_2^2$ converges to one in probability uniformly over $\mathcal{P}_0$ under Assumption 1(a). It can be further shown that the term $(\mathbb{III})/\sigma_2^2$ is $o_{\mathcal{P}_0}(1)$ by Markov's inequality combined with Assumption 1(b). Lastly, the term $(\mathbb{IIII})$ satisfies $(\mathbb{IIII}) \le \{(\mathbb{I}) \times (\mathbb{III})\}^{1/2}$, which is again $o_{\mathcal{P}_0}(1)$. Therefore, $\widehat{\sigma}_2^2/\sigma_2^2$ converges to one in probability uniformly over $\mathcal{P}_0$. This further proves that $\{(\sigma_1^2 + \sigma_2^2)/(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)\}^{1/2} = 1 + o_{\mathcal{P}_0}(1)$ by Lundborg et al. (2022, Lemma S7).

Putting all pieces together, Lemma 4(b), proves the claim.

1.4. *Proof of Corollary 1*

For each $j \in \{1, \ldots, K\}$, the proof of Theorem 3 shows that $(\sigma_1^2 + \sigma_2^2)^{1/2}/(\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2)^{1/2} = 1 + o_{\mathcal{P}_0}(1)$. Thus, by Lemma 4(b), it is enough to show the asymptotic normality of

$$\frac{1}{K^{1/2}} \sum_{j=1}^{K} \frac{m^{1/2}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{(\sigma_1^2 + \sigma_2^2)^{1/2}}.$$

Without loss of generality, denote the sample indices of $D_1, D_2, \ldots, D_K$ as

$$I_1 = \{1, \ldots, m\},\ I_2 = \{m+1, \ldots,\ 2m\}, \ldots, I_K = \{m(K-1)+1, \ldots, mK\}.$$

Then the proof of Theorem 3 establishes that

$$\frac{1}{K^{1/2}} \sum_{j=1}^{K} \frac{m^{1/2}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{(\sigma_1^2 + \sigma_2^2)^{1/2}} = \frac{1}{(mK)^{1/2}} \sum_{j=1}^{K} \underbrace{\left\{ \sum_{i \in I_j} \frac{A_{1,i} + A_{2,i} - 1}{(\sigma_1^2 + \sigma_2^2)^{1/2}} \right\}}_{=B_j} + o_{\mathcal{P}_0}(1),$$

where $A_{1,i} + A_{2,i} - 1 = \mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1$. Note that $B_1, \ldots, B_K$ are mutually independent conditional on $\widehat{h}$. As in the proof of Theorem 3, we apply the conditional central limit theorem (Lemma 3) to the average of $B_1, \ldots, B_K$ conditional on $\widehat{h}$ under Assumption 1, which completes the proof of Corollary 1.

1.5. *Proof of Theorem 4*

The proof consists of two parts as in the proof of Theorem 3. In the first part, we investigate the numerator of $\widehat{\mathrm{MMD}}_\ell^2$, i.e., $\overline{S}$, whereas in the second part, we show the consistency of the denominator to the population variance under Assumption 2. The proof is then completed by Lemma 4(b).

Analysis of the numerator. Using the fact that a kernel can be expressed as an inner product of feature maps, $k(x, y) = \langle \psi(x), \psi(y) \rangle$, we can rewrite $\overline{S}$ as

$$\overline{S} = \frac{1}{m} \sum_{i=1}^{m} \left\langle \psi\left(V_i^{(1)}\right) - \widehat{r}_X\left(X_i^{(2)}\right)\psi\left(V_i^{(2)}\right),\ \psi\left(V_{i+m}^{(1)}\right) - \widehat{r}_X\left(X_{i+m}^{(2)}\right)\psi\left(V_{i+m}^{(2)}\right) \right\rangle.$$

By adding and subtracting $r_X(X_i^{(2)})\psi(V_i^{(2)})$ and $r_X(X_{i+m}^{(2)})\psi(V_{i+m}^{(2)})$, $\overline{S}$ can be written as the sum of the four terms given as:

$$(\mathbb{I}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\left\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_{i+m}^{(1)}) - r_X(X_{i+m}^{(2)})\psi(V_{i+m}^{(2)}) \right\rangle}_{=S_i},$$

$$(\mathbb{II}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\left\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_{i+m}^{(2)}) \right\rangle \left\{\widehat{r}_X(X_{i+m}^{(2)}) - r_X(X_{i+m}^{(2)})\right\}}_{=\widehat{S}_{i,a}},$$

$$(\mathbb{III}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\left\langle \psi(V_i^{(2)}), \psi(V_{i+m}^{(1)}) - r_X(X_{i+m}^{(2)})\psi(V_{i+m}^{(2)}) \right\rangle \left\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\right\}}_{=\widehat{S}_{i,b}},$$

$$(\mathbb{IV}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\left\langle \psi(V_i^{(2)}), \psi(V_{i+m}^{(2)}) \right\rangle \left\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\right\} \left\{\widehat{r}_X(X_{i+m}^{(2)}) - r_X(X_{i+m}^{(2)})\right\}}_{=\widehat{S}_{i,c}}.$$

The first term $(\mathbb{I})$ does not involve an estimate of the density ratio and will be asymptotically Gaussian since it is the sum of i.i.d. random variables under the null hypothesis. The other terms $(\mathbb{II})$, $(\mathbb{III})$, and $(\mathbb{IV})$ are asymptotically negligible under the conditions of the theorem. Hence $\overline{S}$ will be dominated by $(\mathbb{I})$. Let us analyze each term separately.

1. **Term** $(\mathbb{I})$**.** Define $\mathrm{var}_P[S_1] = \sigma_P^2$. Then under Assumption 2(a), Lemma 2 yields

$$\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P(m^{1/2}\sigma_P^{-1}(\mathbb{I}) \le t) - \Phi(t)\right| \to 0.$$

2. **Terms** $(\mathbb{II})$ **and** $(\mathbb{III})$**.** We only analyze the term $(\mathbb{II})$ since $(\mathbb{III})$ can be handled in exactly the same way by symmetry. Under the null hypothesis, by the law of total expectation, it can be seen that the expectation of the summands of $(\mathbb{II})$ is equal to zero:

$$E\left[\left\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_{i+m}^{(2)}) \right\rangle \left\{\widehat{r}_X(X_{i+m}^{(2)}) - r_X(X_{i+m}^{(2)})\right\}\right] = 0,$$

which leads to $E[(\mathbb{II})] = 0$. On the other hand, the conditional second moment (or the conditional variance) of $(\mathbb{II})$ given D_b satisfies

$$\begin{aligned} E\left[(\mathbb{II})^2 \mid D_b\right] &= \frac{1}{m}E\left[\left\langle \psi(V_1^{(1)}) - r_X(X_1^{(2)})\psi(V_1^{(2)}), \psi(V_{1+m}^{(2)}) \right\rangle^2 \left\{\widehat{r}_X(X_{1+m}^{(2)}) - r_X(X_{1+m}^{(2)})\right\}^2 \mid D_b\right] \\ &\lesssim \frac{K^2}{m}\left[1 + E\left\{r_X(X_1^{(2)})^2\right\}\right] E\left[\left\{\widehat{r}_X(X_{1+m}^{(2)}) - r_X(X_{1+m}^{(2)})\right\}^2 \mid D_b\right], \end{aligned}$$

where we use the fact that $\langle \psi(x), \psi(y)\rangle = k(x, y)$, whose ℓ_∞ norm is uniformly bounded by the constant K. Therefore, under the condition that

$$\sup_{P\in\mathcal{P}_0} E_P\left[\left\{\widehat{r}_X(X_{1+m}^{(2)}) - r_X(X_{1+m}^{(2)})\right\}^2\right] = o(m^{-1/2}) \quad \text{and} \quad \sup_{P\in\mathcal{P}_0} E_P\left\{r_X(X_1^{(2)})^2\right\} < \infty,$$

Chebyshev's inequality yields $(\mathbb{II}) = o_{\mathcal{P}_0}(m^{-1/2})$ and similarly $(\mathbb{III}) = o_{\mathcal{P}_0}(m^{-1/2})$.

3. **Term** $(\mathbb{IV})$**.** The fourth term $(\mathbb{IV})$ can be written as

$$(\mathbb{IV}) = \frac{1}{m}\sum_{i=1}^{m} k(V_i^{(2)}, V_{i+m}^{(2)})\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}\ \{\widehat{r}_X(X_{i+m}^{(2)}) - r_X(X_{i+m}^{(2)})\}.$$

Since the kernel is uniformly bounded and by Cauchy–Schwarz inequality, we have

$$|(\mathbb{IV})| \le K\left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}^2\right]^{1/2}\left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{r}_X(X_{i+m}^{(2)}) - r_X(X_{i+m}^{(2)})\}^2\right]^{1/2}$$
$$\overset{(\star)}{=} o_{\mathcal{P}_0}(m^{-1/4})o_{\mathcal{P}_0}(m^{-1/4}) = o_{\mathcal{P}_0}(m^{-1/2}),$$

which follows by Markov's inequality along with the condition that

$$\sup_{P\in\mathcal{P}_0} E_P\left[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2\right] = o(m^{-1/2}),$$

and step $(\star)$ holds by (Lundborg et al., 2022, Lemma S5). Therefore it holds that $(\mathbb{IV}) = o_{\mathcal{P}_0}(m^{-1/2})$.

Now combining the results establishes that

$$\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}} |\mathrm{pr}_P(m^{1/2}\sigma_P^{-1}\overline{S} \le t) - \Phi(t)| \to 0.$$

Consistency of the variance estimate. Denoting

$$\widehat{\sigma}_P^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{S}_i - \overline{S})^2,$$

we would like to show that $\widehat{\sigma}_P^2/\sigma_P^2$ converges to one in probability, which further implies that $\sigma_P/\widehat{\sigma}_P = 1 + o_{\mathcal{P}_0}(1)$ by Lundborg et al. (2022, Lemma S7). Since the test statistic $\widehat{\mathrm{MMD}}_\ell^2$ is scale-invariant, we may assume that $\sigma_P^2 = 1$ without loss of generality. Moreover, the preceding analysis ensures that $\overline{S} = o_{\mathcal{P}_0}(1)$. Therefore we only need to show $\frac{1}{m}\sum_{i=1}^{m}\widehat{S}_i^2$ converges to one in probability.

$$\left|\frac{1}{m}\sum_{i=1}^{m}\widehat{S}_i^2 - 1\right| = \left|\frac{1}{m}\sum_{i=1}^{m}(S_i + \widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c})^2 - 1\right|$$
$$\le \left|\frac{1}{m}\sum_{i=1}^{m}S_i^2 - 1\right| + \left|\frac{1}{m}\sum_{i=1}^{m}(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c})^2\right| + 2\left|\frac{1}{m}\sum_{i=1}^{m}S_i(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c})\right|$$
$$\le \left|\frac{1}{m}\sum_{i=1}^{m}S_i^2 - 1\right| + \left|\frac{1}{m}\sum_{i=1}^{m}(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c})^2\right|$$
$$+ 2\Big(\frac{1}{m}\sum_{i=1}^{m}S_i^2\Big)^{1/2}\Big\{\frac{1}{m}\sum_{i=1}^{m}(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c})^2\Big\}^{1/2},$$

where the last inequality follows by the Cauchy–Schwarz inequality. By the law of large numbers, $\frac{1}{m}\sum_{i=1}^{m}S_i^2$ converges to one in probability. Thus the proof amounts to showing that $\frac{1}{m}\sum_{i=1}^{m}(\widehat{S}_{i,a} +$

$\widehat{S}_{i.b} + \widehat{S}_{i,c}\big)^2 = o_{\mathcal{P}_0}(1)$, which is implied by

$$\frac{1}{m}\sum_{i=1}^{m} \widehat{S}_{i,a}^2 = o_{\mathcal{P}_0}(1), \quad \frac{1}{m}\sum_{i=1}^{m} \widehat{S}_{i,b}^2 = o_{\mathcal{P}_0}(1) \quad \text{and} \quad \frac{1}{m}\sum_{i=1}^{m} \widehat{S}_{i,c}^2 = o_{\mathcal{P}_0}(1).$$

This can be done as the way how (𝕀𝕀), (𝕀𝕀𝕀), and (𝕀𝕍) are handled earlier along with Markov's inequality. This completes the proof.

1.6. *Proof of Corollary 2*

For each $j \in \{1, \ldots, K\}$, the proof of Theorem 4 shows that $\sigma_P/\widehat{\sigma}_j = 1 + o_{\mathcal{P}_0}(1)$. Thus, by Lemma 4(b), it is enough to show the asymptotic normality of

$$\frac{1}{K}\sum_{j=1}^{K} \frac{n^{1/2}\overline{S}_j}{\sigma}.$$

Without loss of generality, denote the sample indices of $D_1, D_2, \ldots, D_K$ as

$$I_1 = \left\{1, \ldots, \frac{2n}{K}\right\}, \; I_2 = \left\{\frac{2n}{K} + 1, \ldots, \; \frac{4n}{K}\right\}, \ldots, I_K = \left\{2n - \frac{2n}{K} + 1, \ldots, 2n\right\},$$

and let the first n/K elements of I_j as I_j' (e.g., $I_1' = \{1, \ldots, n/K\}$). Then with $m' = n/K$, the proof of Theorem 4 establishes that

$$\sum_{j=1}^{K} \overline{S}_j = \frac{K}{n}\sum_{j=1}^{K}\sum_{i \in I_j'} S_i + o_{\mathcal{P}_0}(n^{-1/2}),$$

where

$$S_i = \left\langle \psi\left(V_i^{(1)}\right) - r_X\left(X_i^{(2)}\right)\psi\left(V_i^{(2)}\right), \psi\left(V_{i+m'}^{(1)}\right) - r_X\left(X_{i+m'}^{(2)}\right)\psi\left(V_{i+m'}^{(2)}\right)\right\rangle.$$

Notably, $\sum_{i \in I_1'} S_i, \ldots, \sum_{i \in I_K'} S_i$ are mutually independent. Hence

$$\operatorname{var}\left(\frac{K}{n}\sum_{j=1}^{K}\sum_{i \in I_j'} \sigma_P^{-1} S_i\right) = \frac{K^2}{n^2}\, K\, \frac{n}{K} = \frac{K^2}{n}.$$

By the central limit theorem (Lemma 2),

$$\frac{1}{K}\sum_{j=1}^{K}\sum_{i \in I_j'} \frac{n^{1/2} S_i}{\sigma_P} = \frac{1}{K}\sum_{j=1}^{K} \frac{n^{1/2}\overline{S}_j}{\sigma} + o_{\mathcal{P}_0}(1)$$

converges to $N(0, 1)$ as desired.

1.7. *Proof of Example 1*

In this section, we aim to present a detailed analysis of the asymptotic equivalence between the GCM statistic T and its counterpart $\tilde{T}$ constructed using $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$. Throughout this section, we assume that Y has a finite second moment and that $Y \perp\!\!\!\perp Z \mid X$, i.e., the null hypothesis holds. Let

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

and $\tilde{T} = \nu_{\tilde{R}}/\hat{\sigma}_{\tilde{R}}$. Let $\sigma^2_{\tilde{R}} > 0$ denote the variance of $\{\tilde{Y} - f(\tilde{X})\}\{\tilde{Z} - g(\tilde{X})\}$ where $(\tilde{X}, \tilde{Y}, \tilde{Z})$ is a random draw from the joint distribution P_{XYZ}. We begin with an upper bound for $|T - \tilde{T}|$:

$$\begin{aligned}\left|\frac{\nu_R}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_{\tilde{R}}}\right| &\leq \left|\frac{\nu_R}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_R}\right| + \left|\frac{\nu_{\tilde{R}}}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_{\tilde{R}}}\right| \\ &\leq \frac{1}{\hat{\sigma}_R}|\nu_R - \nu_{\tilde{R}}| + \frac{|\nu_{\tilde{R}}|}{(\hat{\sigma}_R + \hat{\sigma}_{\tilde{R}})\hat{\sigma}_R\hat{\sigma}_{\tilde{R}}}\left|\hat{\sigma}^2_R - \hat{\sigma}^2_{\tilde{R}}\right|,\end{aligned}$$

from which the proof boils down to showing the convergence of the following four terms to zero in probability: (a) $\nu_R - \nu_{\tilde{R}}$, (b) $\hat{\sigma}^2_R - \hat{\sigma}^2_{\tilde{R}}$, (c) $\hat{\sigma}^2_R - \sigma^2_{\tilde{R}}$ and (d) $\hat{\sigma}^2_{\tilde{R}} - \sigma^2_{\tilde{R}}$. Under these convergence results, the asymptotic equivalence follows by the continuous mapping theorem along with the fact that $\nu_{\tilde{R}}$ is stochastically bounded by the central limit theorem. In what follows, we establish convergence of (a), (b), (c), and (d) to zero in probability in order.

1. **Term (a): $\nu_R - \nu_{\tilde{R}}$.** Starting with the term (a), the difference between ν_R and $\nu_{\tilde{R}}$ can be written as

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n} R_i - \frac{1}{n^{1/2}}\sum_{i=1}^{n}\tilde{R}_i = \underbrace{\frac{1}{n^{1/2}}\sum_{i=n_1+1}^{\bar{n}_1}(R_i - \tilde{R}_i)\,\mathbb{1}(\bar{n}_1 > n_1)}_{=\Delta_1} + \underbrace{\frac{1}{n^{1/2}}\sum_{i=\bar{n}_1+1}^{n_1}(R_i - \tilde{R}_i)\,\mathbb{1}(\bar{n}_1 \leq n_1)}_{=\Delta_2}. \tag{4}$$

   Remark that Z_i is a fixed constant for a given index i, which allows us to show that $E(R_i) = 0$ for any $i \in [n]$. For example, when $i = 1$, Z_1 equals 1 (since $X_1 = X_1^{(1)}$) and thus the law of total expectation yields

$$E(R_1) = E[\{1 - g(X_1)\}\{Y_1 - f(X_1)\}] = E\left[\{1 - g(X_1)\}\,E[\{Y_1 - f(X_1)\} \mid X_1]\right] = 0,$$

   where we recall $f(X_1) = E(Y_1 \mid X_1)$. It also follows that $E(\tilde{R}_i) = 0$ for any $i \in [n]$ under the null hypothesis. This together with the law of total expectation shows that

$$E[\Delta_1] = E\left\{\frac{1}{n^{1/2}}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{1}(\bar{n}_1 > n_1)E\left(R_i - \tilde{R}_i \mid \bar{n}_1\right)\right\} = 0,$$

   and similarly $E(\Delta_2) = 0$. Thus, we have $E(\nu_R - \nu_{\tilde{R}}) = 0$.
   Now consider the variance of $\nu_R - \nu_{\tilde{R}}$. Since $E(\nu_R - \nu_{\tilde{R}}) = 0$ and $E(\Delta_1\Delta_2) = 0$, we have

$$\mathrm{var}(\nu_R - \nu_{\tilde{R}}) = \mathrm{var}(\Delta_1) + \mathrm{var}(\Delta_2).$$

For $\mathrm{var}(\Delta_1)$, we have

$$\begin{aligned}
\mathrm{var}(\Delta_1) &= E\big\{\mathrm{var}\big(\Delta_1 \mid \bar{n}_1\big)\big\} + \mathrm{var}\big\{\underbrace{E\big(\Delta_1 \mid \bar{n}_1\big)}_{=0}\big\} \\
&= E\left[\mathrm{var}\left\{\frac{1}{n^{1/2}}\sum_{i=n_1+1}^{\bar{n}_1}(R_i - \tilde{R}_i)\,\mathbb{1}(\bar{n}_1 > n_1)\;\Big|\;\bar{n}_1\right\}\right] \\
&= E\left[\frac{1}{n}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{1}(\bar{n}_1 > n_1)\,E\big\{\big(R_i - \tilde{R}_i\big)^2 \mid \bar{n}_1\big\}\right] \\
&\overset{\text{(i)}}{\leq} 2E\left\{\frac{\bar{n}_1 - n_1}{n}\,\frac{1}{\bar{n}_1 - n_1}\sum_{i=n_1+1}^{\bar{n}_1}E\big(R_i^2 + \tilde{R}_i^2 \mid \bar{n}_1\big)\right\} \overset{\text{(ii)}}{\leq} \frac{4}{n}E\left(|\bar{n}_1 - n_1|\right)\mathrm{var}(Y_1) \overset{\text{(iii)}}{\leq} \frac{2}{n^{1/2}}\mathrm{var}(Y_1),
\end{aligned}$$

where (i) follows from the inequality $(x-y)^2 \leq 2x^2 + 2y^2$ and (ii) uses the law of total variance along with the fact that $R_i^2 \leq \{Y_i - f(X_i)\}^2$ and $\tilde{R}_i^2 \leq \{\tilde{Y}_i - f(\tilde{X}_i)\}^2$ since $Z_i, \tilde{Z}_i \in \{1,2\}$. For the last inequality (iii), we use $E(|\bar{n}_1 - n_1|) \leq n^{1/2}/2$. The same bound holds for $\mathrm{var}(\Delta_2)$ and thus

$$\mathrm{var}(\nu_R - \nu_{\tilde{R}}) \leq \frac{4}{n^{1/2}}\mathrm{var}(Y_1).$$

Combining the results with Chebyshev's inequality now shows that $\nu_R - \nu_{\tilde{R}}$ converges to zero in probability.

2. **Term (b): $\hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2$.** We next aim to show that

$$\hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2 = \left\{\frac{1}{n}\sum_{i=1}^n R_i^2 - \Big(\frac{1}{n}\sum_{r=1}^n R_r\Big)^2\right\} - \left\{\frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 - \Big(\frac{1}{n}\sum_{r=1}^n \tilde{R}_r\Big)^2\right\}$$

converges to zero in probability. We decompose this into two terms

$$(\mathbb{I}) = \frac{1}{n}\sum_{i=1}^n R_i^2 - \frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 \quad \text{and} \quad (\mathbb{II}) = \Big(\frac{1}{n}\sum_{r=1}^n R_r\Big)^2 - \Big(\frac{1}{n}\sum_{r=1}^n \tilde{R}_r\Big)^2,$$

and show each of them converges to zero in probability. For the first term $(\mathbb{I})$, we have

$$\frac{1}{n}\sum_{i=1}^n R_i^2 - \frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 = \underbrace{\frac{1}{n}\sum_{n_1+1}^{\bar{n}_1}(R_i^2 - \tilde{R}_i^2)\,\mathbb{1}(\bar{n}_1 > n_1)}_{=\tilde{\Delta}_1} + \underbrace{\frac{1}{n}\sum_{\bar{n}_1+1}^{n_1}(R_i^2 - \tilde{R}_i^2)\,\mathbb{1}(\bar{n}_1 \leq n_1)}_{=\tilde{\Delta}_2}.$$

Using the law of total expectation, we obtain

$$\begin{aligned}
E(|\tilde{\Delta}_1|) &= E\left[\Big|\frac{1}{n}\sum_{i=n_1+1}^{\bar{n}_1}(R_i^2 - \tilde{R}_i^2)\,\mathbb{1}(\bar{n}_1 > n_1)\Big|\right] \\
&\leq E\left\{\frac{\bar{n}_1 - n_1}{n}\,\frac{1}{\bar{n}_1 - n_1}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{1}(\bar{n}_1 > n_1)\,E\big(|R_i^2 - \tilde{R}_i^2| \mid \bar{n}_1\big)\right\} \\
&\leq \frac{2}{n}E(|\bar{n}_1 - n_1|)\mathrm{var}(Y_1) \leq \frac{\mathrm{var}(Y_1)}{n^{1/2}},
\end{aligned}$$

where the first inequality is derived from Jensen's inequality, and the remaining steps follow from the previous results. A similar argument applies to $\tilde{\Delta}_2$. By Markov's inequality, (𝕀) converges to zero. For the second term (𝕀𝕀), we have

$$\left|\left(\frac{1}{n}\sum_{r=1}^{n} R_r\right)^2 - \left(\frac{1}{n}\sum_{r=1}^{n} \tilde{R}_r\right)^2\right| \leq \left|\frac{1}{n}\sum_{r=1}^{n} R_r + \frac{1}{n}\sum_{r=1}^{n} \tilde{R}_r\right| \left|\frac{1}{n}\sum_{r=1}^{n} R_r - \frac{1}{n}\sum_{r=1}^{n} \tilde{R}_r\right|,$$

which can be shown to converge to zero in probability using the previous results.

3. **Terms (c) and (d): $\hat{\sigma}_R^2 - \sigma_{\tilde{R}}^2$ and $\hat{\sigma}_{\tilde{R}}^2 - \sigma_{\tilde{R}}^2$.** We can see that the term (d) converges to zero in probability by the conventional law of large numbers. The term (c) also converges to zero as well since $\hat{\sigma}_R^2 - \sigma_{\tilde{R}}^2 = (\mathrm{c})' + (\mathrm{c})''$, where $(\mathrm{c})' = \hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2$ and $(\mathrm{c})'' = \hat{\sigma}_{\tilde{R}}^2 - \sigma_{\tilde{R}}^2$. In fact, $(\mathrm{c})' = (\mathrm{b})$ and $(\mathrm{c})'' = (\mathrm{d})$, and both are known to converge to zero in probability based on the previous results.

This completes the proof which shows that T and $\tilde{T}$ are asymptotically equivalent for the stable case.

2. Supporting Lemmas

In this section, we collect several lemmas from the existing literature for completeness. The proof of the following lemma can be found, for example, in Mulzer (2018).

Lemma 1. *Let $Z_1, \ldots, Z_n$ be i.i.d. Bernoulli random variables with success probability $p \in [0,1]$ and $S_n = \sum_{i=1}^n Z_i$. For any $\delta \in [0,1]$, it holds that*

$$\mathrm{pr}\{S_n \geq (1+\delta)np\} \leq e^{-\frac{np\delta^2}{3}} \quad \text{and} \quad \mathrm{pr}\{S_n \leq (1-\delta)np\} \leq e^{-\frac{np\delta^2}{3}}.$$

The following is the uniform central limit theorem result in Shah & Peters (2020, Lemma 18).

Lemma 2. *(Shah & Peters, 2020, Lemma 18) Let $\mathcal{P}$ be a family of distributions for a random variable $\zeta \in \mathbb{R}$ and suppose that $\zeta_1\zeta_2, \ldots$ are i.i.d. copies of ζ. For each $n \in \mathbb{N}$, let $S_n = n^{-1/2}\sum_{i=1}^n \zeta_i$. Suppose that for all $P \in \mathcal{P}$, we have $E_P(\zeta) = 0$, $E_P(\zeta^2) = 1$ and $E_P(|\zeta|^{2+\eta}) < c$ for some $\eta, c > 0$. We have that*

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

The next lemma corresponds to Lundborg et al. (2022, Lemma S8) on conditional uniform central limit theorem.

Lemma 3. *(Lundborg et al., 2022, Lemma S8) Let $(X_{n,i})_{n\in\mathbb{N}, i\in[n]}$ be a triangular array of real-valued random variables and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration on $\mathcal{F}$. Assume that*

1. *$X_{n,1}, \ldots, X_{n,n}$ are conditionally independent given $\mathcal{F}_n$, for each $n \in \mathbb{N}$;*
2. *$E_P(X_{n,i} \mid \mathcal{F}_n) = 0$ for all $n \in \mathbb{N}, i \in [n]$;*
3. *$\left|n^{-1}\sum_{i=1}^n E_P(X_{n,i}^2 \mid \mathcal{F}_n) - 1\right| = o_{\mathcal{P}}(1)$;*
4. *there exists $\delta > 0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} E_P\left(|X_{n,i}|^{2+\delta} \mid \mathcal{F}_n\right) = o_{\mathcal{P}}(n^{\delta/2}).$$

Then $S_n = n^{-1/2} \sum_{m=1}^{n} X_{n,m}$ *converges uniformly in distribution to* $N(0,1)$*, i.e.*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{x\in\mathbb{R}} |\mathrm{pr}_P(S_n \le x) - \Phi(x)| = 0.$$

The next lemma corresponds to Lundborg et al. (2022, Lemma 20) on uniform Slutsky's theorem.

Lemma 4. *(Shah & Peters, 2020, Lemma 20) Let* $\mathcal{P}$ *be a family of distributions that determines the law of a sequences* $(V_n)_{n\in\mathbb{N}}$ *and* $(W_n)_{n\in\mathbb{N}}$ *of random variables. Suppose*

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

Then we have the following.

(a) If $W_n = o_{\mathcal{P}}(1)$*, we have* $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P\left(V_n + W_n \le t\right) - \Phi(t)\right| = 0$*.*
(b) If $W_n = 1 + o_{\mathcal{P}}(1)$*, we have* $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} \left|\mathrm{pr}_P\left(V_n/W_n \le t\right) - \Phi(t)\right| = 0$*.*

3. Additional Numerical Experiments

In this section, we present the experimental details and additional numerical experiments not included in the main body of the paper. Specifically, we describe the experimental details in Appendicies 3.1 and 3.2 and provide additional simulation results, including an empirical analysis of the impact of density ratio estimation errors on DRT methods in Section 3.3, a real data analysis for CIT methods in Section 3.4, a sensitivity analysis of the application of Algorithm 1 to the CIT approach in Section 3.5, and a sensitivity analysis of Algorithm 1 to adjustment parameter ε in Section 3.6.

3.1. *Overview of Testing Methods*

This subsection outlines the testing methods employed in our experiments. We denote the single-split classifier-based test in the main part as CLF and its cross-fit version as †CLF. Both classifier-based test statistics are built on a specific form of classifiers detailed in Section 3.2. Moreover, we denote the linear-time MMD in the main part as MMD-ℓ and its cross-fit version as †MMD-ℓ. Additional conditional two-sample testing methods included in our experiments are as follows:

CP: The conformal prediction (CP) test utilizes a conformity score to produce a weighted rank sum test statistic. This statistic is constructed by estimating both marginal and conditional density ratios, which can be approached using various density ratio estimation methods. For further details, please refer to Hu & Lei (2024).

DCP: The debiased conformal prediction (DCP) test refines the CP test by reducing bias through the use of Neyman orthogonality and using cross-fitting to improve efficiency. This enhancement guarantees asymptotic normality under certain conditions. Further technical details and theoretical guarantees are described in Chen & Lei (2024).

For CIT methods, we employ one kernel-based and three regression-based testing approaches. All of these CIT methods are implemented via Algorithm 1. We empirically observe in Section 3.5

that the performance of the CIT methods remains largely consistent regardless of whether Algorithm 1 is applied or not, especially when the sample size is large. The following methods are included in our experiments:

RCIT: The randomized conditional independence test (RCIT) approximates the kernel conditional independence test by leveraging random Fourier features, allowing it to scale linearly with sample size. We use the default options in its implementation, and further details are provided in Strobl et al. (2019).

GCM: The generalized covariance measure (GCM) by Shah & Peters (2020) utilizes the normalized covariance between residuals from regression models as a test statistic. This approach provides a flexible framework that can be adapted to various settings by selecting appropriate regression techniques.

PCM: The projected covariance measure (PCM) is a variation of the GCM applied to a transformed version of X. For our simulations, we follow Algorithm 1 from Lundborg et al. (2022). This method retains the general structure of the GCM while introducing a projection step, which enhances power, particularly when the conditional covariance is zero or near zero.

WGSC: This testing procedure proposes a general framework for nonparametric inference on interpretable, algorithm-agnostic variable importance. In our simulation, we follow the approach outlined in Williamson et al. (2023, Algorithm 3), which utilizes sample splitting and cross-fitting.

3.2. *Experimental Details*

We begin with the implementation details of our numerical experiments, including density ratio estimation techniques, linear-time MMD test, classifier-based test, and conditional independence testing approaches.

Density Ratio Estimation. We estimate the density ratio $r_X(x)$ defined in the main part using a probabilistic classification-based approach described in Sugiyama et al. (2010, Section 3). Specifically, we focus on two classifiers: linear logistic regression (LL) and kernel logistic regression (KLR).

Given samples $\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(1)}$ and $\{(X_j^{(2)}, Y_j^{(2)})\}_{j=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(2)}$, consider $\{(X_i, \ell_i)\}_{i=1}^{n}$, where $(X_1, \ldots, X_n) = (X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ and $(\ell_1, \ldots, \ell_{n_1}, \ell_{n_1+1}, \ldots, \ell_n) = (0, \ldots, 0, 1, \ldots, 1)$ with $\ell_i = \mathbb{1}(i \geq n_1 + 1)$. Further denote $X_i = \big(X_i(1), \ldots, X_i(p)\big)^\top$ where p is the dimension of X_i and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

- For LL method, we model the posterior probability as

$$\eta(X_i; \boldsymbol{\beta}) = \mathrm{pr}(\ell = 1 \mid X_i) = \frac{1}{1 + \exp\left(-\beta_0 + \sum_{j=1}^{p} \beta_j X_i(j)\right)}.$$

  The estimated coefficients $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the negative log-likelihood.
- For KLR method (Zhu & Hastie, 2005), we use $\eta(X_i; \boldsymbol{\beta}) = 1/(1 + \exp(-\theta(X_i; \boldsymbol{\beta})))$, where $\theta(X_i; \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p} \beta_j k(X_i(j), x)$ and $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$. The estimated co-

efficients $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the following penalized negative log-likelihood:

$$-\sum_{i=1}^{n}\left[\ell_i\theta(X_i;\boldsymbol{\beta}) - \log\left(1+\exp\left(\theta(X_i;\boldsymbol{\beta})\right)\right)\right] + \frac{\lambda}{2}\|\theta\|^2_{\mathcal{H}_k},$$

where $\mathcal{H}_k$ is the reproducing kernel Hilbert space generated by k and λ is a regularization parameter.

The density ratio estimate is then:

$$\hat{r}_X(X_i) = \frac{n_2}{n_1}\,\frac{\eta(X_i;\hat{\boldsymbol{\beta}})}{1-\eta(X_i;\hat{\boldsymbol{\beta}})}. \tag{3}$$

For the joint density ratio, we use (X_i, Y_i) instead of X_i alone. We set $\sigma^2 = 200$, following Hu & Lei (2024), and fix $\lambda = 0.0005$ throughout our simulations.

Linear-Time MMD Test. For the linear-time MMD tests, we use a Gaussian kernel with the bandwidth parameter fixed at 1 across all experiments. In the cross-validated version (†MMD-ℓ), we use 2-fold cross validation (i.e., $K = 2$) with an equal splitting ratio.

Classifier-based Test. As mentioned in the main part, under the balanced-sample setting, the Bayes optimal classifier is defined as:

$$h^*(y,x) = \mathbb{1}\left(\frac{f^{(1)}_{YX}(y,x)}{f^{(1)}_{YX}(y,x)+f_{YX}(y,x)} > \frac{1}{2}\right).$$

This classifier can be equivalently expressed using density ratios:

$$h^\star(y,x) = \mathbb{1}\left(\frac{r_X(x)}{r_X(x)+r_{YX}(y,x)} > \frac{1}{2}\right),$$

where $r_X(x) = f^{(1)}_X(x)/f^{(2)}_X(x)$ as in the main part and $r_{YX}(y,x) = f^{(1)}_{YX}(y,x)/f^{(2)}_{YX}(y,x)$. The empirical classifier is then defined as a plug-in estimator of $h^\star$:

$$\hat{h}(y,x) = \mathbb{1}\left(\frac{\hat{r}_X(x)}{\hat{r}_X(x)+\hat{r}_{YX}(y,x)} > \frac{1}{2}\right), \tag{4}$$

where $\hat{r}_X$ and $\hat{r}_{YX}$ are the estimated marginal and joint density ratios, respectively, obtained using the classification-based approach described above.

The classifier $\hat{h}$ is constructed based on the training set, whereas the testing set is further split into two subsets with a ratio of 8:2. A larger subset is used for density ratio estimation, and the other subset is used for calculating the test statistic. In the cross-validated version (†CLF), we use 2-fold cross-validation as in †MMD-ℓ, maintaining 8:2 splitting ratio within each fold.

Randomized Conditional Independence Test. The RCIT method is implemented using the default hyperparameter settings specified in Strobl et al. (2019). Specifically, we use the default approximation method (Lindsay–Pilla–Basak method) for the null distribution. The number of random Fourier features is set to 100 for the conditioning set, and 5 for the non-conditioning sets.

Regression Methods for CIT. In our implementation of CIT approaches, we utilize several standard regression techniques. Table 1 below provides an overview of the key methods and their corresponding hyperparameter settings.

Table 1: Description of regression methods used in CIT approach.

Regression Method	R Implementation	Tuning Parameters	Description
Random Forests	`ranger`	$mtry = p^{1/2}$	# of variables to split at each node
XGBoost	`xgboost`	max depth = 6 $\eta = 0.3$	maximum tree depth learning rate

The code for reproducing all of our simulation results (including those in Sections 3.3 to 3.6) and for more detailed settings is available on GitHub: `https://github.com/suman-cha/Cond2ST`.

3.3. *Impact of Density Ratio Estimation Errors on DRT Methods*

To complement the real data analysis presented in the main part, we conduct experiments to examine the relationship between density ratio estimation errors and type I errors of DRT methods. Fig. 1 illustrates the log-scaled mean squared error (MSE) of the marginal density ratio r_X and the conditional density ratio $r_{Y|X}$ estimates for both the LL and KLR methods across various sample sizes. Our experimental setup involves 500 simulations for each combination of sample size, dataset, and estimation method. We report the median MSE to provide a robust measure of estimation accuracy. For better visualization, the log-scaled MSE values are clipped: marginal density ratio errors above 10 are capped at 10 and conditional density ratio errors are limited to a maximum of 1. Notably, the true error values for the LL method in high-dimensional settings, significantly exceed these clipped limits.

In the low-dimensional diamonds dataset, both LL and KLR methods show relatively low MSE values for both marginal and conditional density ratio estimation. The performance gap between LL and KLR methods diminishes as the sample size increases. This observation aligns with the findings in the main part, where simpler methods like LL suffice to control type I error in low-dimensional settings. The low estimation errors explain their similar performance in such scenarios.

In contrast, the high-dimensional superconductivity dataset shows significant differences between the methods. KLR consistently outperforms LL in both marginal and conditional density ratio estimation, maintaining low and stable MSE values across all sample sizes. On the other hand, LL shows extremely high MSE values, particularly for small sample sizes. Although LL shows some improvement as the sample size increases, it remains inferior to KLR in terms of estimation accuracy.

The high estimation errors for LL in high-dimensional settings, even beyond the clipping applied, account for poor type I error control observed in the main part. These results highlight the need for more advanced density ratio estimation techniques to ensure the validity of tests in complex and high-dimensional scenarios.

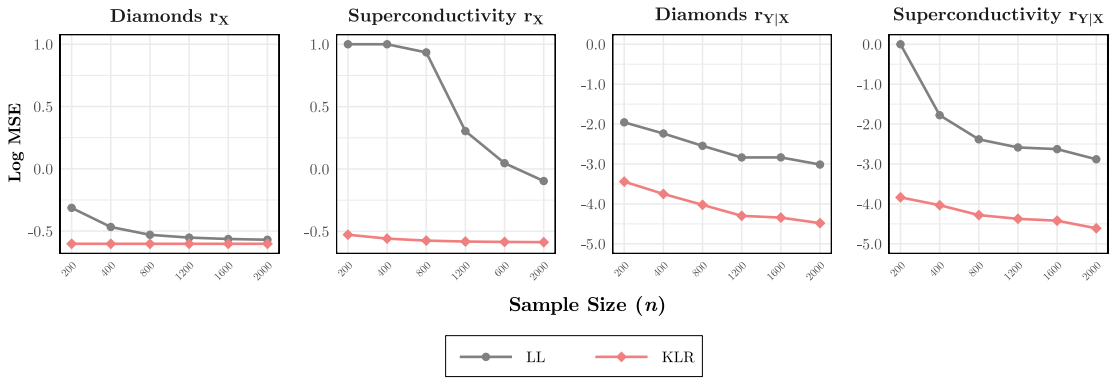

Fig. 1: Log-scaled mean squared errors of marginal density ratio $r_X(x)$ (*left*) and conditional density ratio $r_{Y|X}(y|x)$ (*right*) estimates for LL and KLR methods across various sample sizes. Results are shown for diamonds and superconductivity datasets, based on median values from 500 simulations under the null hypothesis.

3.4. *Real Data Analysis for CIT Methods*

We present the results for CIT methods applied to the diamonds and superconductivity datasets, complementing the analysis discussed in main paper. Fig. 2 shows the rejection rates for these methods under both the null and alternative hypotheses across various sample sizes.

For the low-dimensional diamonds dataset, the CIT methods generally exhibit good type I error control, with rejection rates close to the significance level $\alpha = 0.05$ under the null hypothesis. Under the alternative hypothesis, we observe increasing power for all methods except for WGSC as the sample size grows. Notably, RCIT and GCM show superior performance in terms of power.

In the high-dimensional superconductivity dataset, the performance of CIT methods is similar to that observed in the diamonds dataset, with no significant differences compared to DRT methods as shown in the main part. In terms of type I error control, GCM exhibits increasing rejection rates under the null hypothesis as the sample size grows. RCIT shows more inflated type I error, especially at small sample size. Regarding power, the CIT methods demonstrate relatively consistent performance across both datasets.

3.5. *Sensitivity Analysis of CIT Methods to Algorithm 1*

We examine the impact of Algorithm 1 on CIT methods across scenarios with unbounded marginal density ratios, as outlined in the main part. Our analysis encompasses RCIT and the regression-based methods PCM, GCM, and WGSC. The latter three are implemented using various regression models, such as linear models (`lm`), Random Forests (`rf`), and XGBoost (`xgb`). Tables 2 to 4 show the results for Scenarios 1(U), 2(U), and 3(U), respectively. In these tables, a checkmark (✓) indicates that Algorithm 1 is applied, whereas a cross (×) indicates it is not.

RCIT exhibits significant sensitivity to the application of Algorithm 1, particularly in Scenario 1(U). In this case, without the algorithm, the type I error rates of RCIT increase with sample size, whereas with the algorithm, these rates decrease as the sample size grows. This behavior highlights the potential stabilizing effect of Algorithm 1 on the performance of RCIT. On the other hand, GCM shows consistent performance across different regression methods, suggesting the robustness to the choice of underlying regression models. In contrast, the performance of PCM

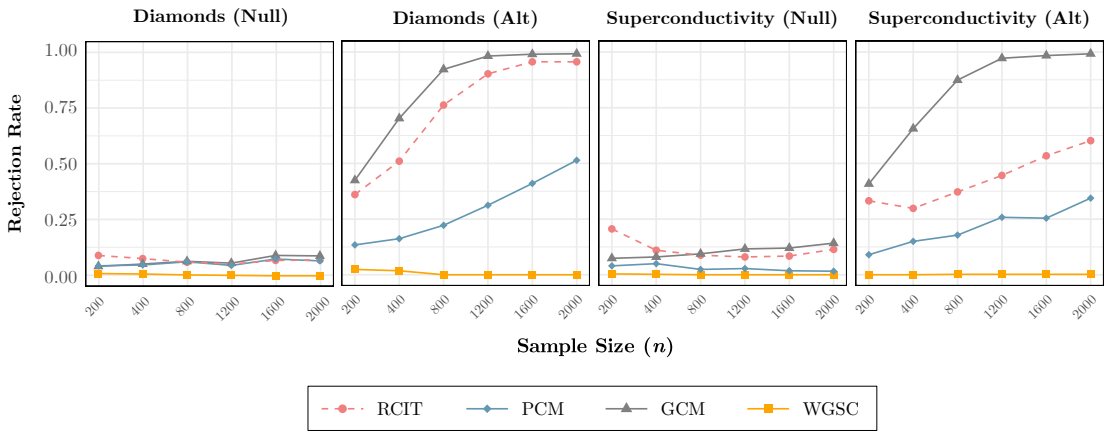

Fig. 2: Rejection rates of CIT methods on the diamonds and superconductivity datasets under null and alternative hypotheses. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

varies significantly with the choice of regression method. WGSC shows inconsistent patterns across scenarios, indicating potential sensitivity to specific data properties or model assumptions. This variability underscores the need for careful consideration when applying WGSC for conditional two-sample testing.

3.6. *Sensitivity Analysis of Algorithm 1 to ε*

We conduct a sensitivity analysis of Algorithm 1 with respect to the adjustment parameter ε. This parameter determines the size of the constructed testing dataset $\mathcal{D}_{\tilde{n}}$, through the equation $\tilde{n} = kn$, where k is a function of ε, as defined in Algorithm 1. The goal of this section is to examine the impact of different ε values on the performance of conditional two-sample testing via CIT methods. We consider 3 candidates for ε: $\{1/n, 1/\log(n), 1/\sqrt{\log(n)}\}$. The analysis covers scenarios with unbounded marginal density ratios, as described in the main part, examining both null and alternative hypotheses across different sample sizes. The settings of these experiments correspond to the settings detailed in the main part with $n \in \{200, 500, 1000\}$.

Tables 5 to 7 present the results for each scenario, comparing the performance of RCIT, PCM, GCM, and WGSC across different ε values and sample sizes. Although the performance of the CIT methods remains relatively stable across different choices of ε, some variations in rejection rates are observed, particularly for small sample sizes.

Table 2: Performance comparison of CIT methods for Scenario 1(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.166	0.034	0.060	0.104	0.048	0.028	0.082	0.000	0.278	0.076
		×	0.202	0.044	0.064	0.098	0.032	0.036	0.074	0.000	0.062	0.052
	Alternative	✓	0.240	0.038	0.068	0.102	0.164	0.076	0.144	0.000	0.284	0.082
		×	0.362	0.046	0.078	0.104	0.300	0.178	0.268	0.000	0.082	0.074
500	Null	✓	0.056	0.066	0.066	0.074	0.036	0.024	0.060	0.000	0.248	0.050
		×	0.210	0.040	0.060	0.078	0.046	0.050	0.064	0.000	0.070	0.076
	Alternative	✓	0.228	0.070	0.092	0.100	0.356	0.170	0.296	0.000	0.252	0.048
		×	0.536	0.038	0.138	0.126	0.552	0.444	0.602	0.000	0.086	0.078
1000	Null	✓	0.064	0.050	0.050	0.060	0.048	0.024	0.060	0.000	0.270	0.056
		×	0.296	0.056	0.038	0.068	0.044	0.062	0.058	0.000	0.068	0.026
	Alternative	✓	0.394	0.050	0.098	0.124	0.586	0.262	0.516	0.000	0.268	0.054
		×	0.762	0.054	0.178	0.182	0.764	0.650	0.858	0.000	0.054	0.044
2000	Null	✓	0.048	0.054	0.030	0.056	0.046	0.028	0.062	0.000	0.282	0.016
		×	0.500	0.036	0.042	0.056	0.048	0.054	0.048	0.000	0.038	0.022
	Alternative	✓	0.610	0.050	0.112	0.110	0.796	0.502	0.820	0.000	0.282	0.024
		×	0.914	0.036	0.292	0.272	0.890	0.812	0.998	0.000	0.086	0.028

Table 3: Performance comparison of CIT methods for Scenario 2(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.162	0.036	0.074	0.098	0.072	0.044	0.084	0.000	0.256	0.060
		×	0.116	0.046	0.048	0.090	0.048	0.058	0.110	0.000	0.056	0.040
	Alternative	✓	0.730	0.038	0.466	0.188	0.064	0.044	0.104	0.000	0.284	0.192
		×	0.808	0.050	0.504	0.254	0.054	0.050	0.110	0.000	0.200	0.174
500	Null	✓	0.074	0.056	0.060	0.104	0.044	0.046	0.094	0.000	0.270	0.042
		×	0.094	0.042	0.042	0.076	0.060	0.056	0.072	0.000	0.058	0.056
	Alternative	✓	0.906	0.056	0.864	0.564	0.054	0.050	0.092	0.000	0.320	0.330
		×	0.880	0.040	0.924	0.702	0.044	0.048	0.074	0.000	0.364	0.494
1000	Null	✓	0.086	0.046	0.032	0.070	0.048	0.020	0.056	0.000	0.272	0.030
		×	0.070	0.048	0.040	0.094	0.048	0.044	0.094	0.000	0.046	0.040
	Alternative	✓	0.968	0.046	0.990	0.974	0.052	0.026	0.066	0.000	0.398	0.604
		×	0.954	0.048	1.000	0.998	0.040	0.048	0.080	0.000	0.724	0.832
2000	Null	✓	0.048	0.036	0.030	0.064	0.032	0.018	0.048	0.000	0.272	0.030
		×	0.076	0.040	0.036	0.070	0.042	0.046	0.072	0.000	0.046	0.020
	Alternative	✓	0.980	0.034	1.000	1.000	0.030	0.020	0.044	0.000	0.486	0.900
		×	0.980	0.042	1.000	1.000	0.040	0.026	0.066	0.000	0.960	0.994

Table 4: Performance comparison of CIT methods for Scenario 3(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.166	0.048	0.062	0.100	0.052	0.028	0.096	0.000	0.254	0.064
		×	0.142	0.042	0.048	0.086	0.058	0.056	0.098	0.000	0.066	0.056
	Alternative	✓	0.688	0.046	0.726	0.578	0.214	0.204	0.266	0.000	0.406	0.636
		×	0.718	0.040	0.758	0.646	0.236	0.248	0.282	0.000	0.658	0.670
500	Null	✓	0.076	0.058	0.064	0.072	0.046	0.030	0.088	0.000	0.268	0.038
		×	0.066	0.042	0.034	0.070	0.046	0.050	0.080	0.000	0.056	0.060
	Alternative	✓	0.734	0.062	0.804	0.762	0.268	0.250	0.294	0.000	0.426	0.708
		×	0.764	0.046	0.808	0.794	0.264	0.270	0.292	0.000	0.742	0.754
1000	Null	✓	0.064	0.052	0.034	0.072	0.050	0.034	0.066	0.000	0.272	0.020
		×	0.086	0.048	0.046	0.072	0.036	0.036	0.080	0.000	0.052	0.036
	Alternative	✓	0.774	0.056	0.826	0.832	0.244	0.214	0.246	0.000	0.470	0.762
		×	0.818	0.048	0.830	0.838	0.236	0.236	0.268	0.000	0.800	0.806
2000	Null	✓	0.070	0.034	0.024	0.062	0.078	0.050	0.084	0.000	0.274	0.026
		×	0.082	0.046	0.036	0.084	0.076	0.068	0.098	0.000	0.044	0.024
	Alternative	✓	0.790	0.038	0.806	0.816	0.224	0.220	0.250	0.000	0.472	0.790
		×	0.810	0.044	0.814	0.824	0.232	0.232	0.268	0.000	0.816	0.802

Table 5: Sensitivity analysis of Algorithm 1 for Scenario 1(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.164	0.064	0.022	0.072
		$1/\log(n)$	0.166	0.064	0.036	0.076
		$1/\sqrt{\log(n)}$	0.168	0.072	0.018	0.062
	Alternative	$1/n$	0.220	0.070	0.080	0.068
		$1/\log(n)$	0.240	0.068	0.076	0.082
		$1/\sqrt{\log(n)}$	0.224	0.074	0.072	0.066
500	Null	$1/n$	0.090	0.038	0.026	0.052
		$1/\log(n)$	0.056	0.066	0.024	0.050
		$1/\sqrt{\log(n)}$	0.080	0.036	0.036	0.044
	Alternative	$1/n$	0.250	0.050	0.126	0.040
		$1/\log(n)$	0.228	0.092	0.170	0.048
		$1/\sqrt{\log(n)}$	0.244	0.062	0.166	0.046
1000	Null	$1/n$	0.046	0.052	0.030	0.038
		$1/\log(n)$	0.064	0.050	0.024	0.056
		$1/\sqrt{\log(n)}$	0.062	0.040	0.030	0.040
	Alternative	$1/n$	0.376	0.078	0.258	0.048
		$1/\log(n)$	0.394	0.098	0.262	0.054
		$1/\sqrt{\log(n)}$	0.394	0.078	0.288	0.038

Table 6: Sensitivity analysis of Algorithm 1 for Scenario 2(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.168	0.070	0.030	0.064
		$1/\log(n)$	0.162	0.074	0.044	0.060
		$1/\sqrt{\log(n)}$	0.164	0.068	0.028	0.064
	Alternative	$1/n$	0.694	0.418	0.026	0.200
		$1/\log(n)$	0.730	0.466	0.044	0.192
		$1/\sqrt{\log(n)}$	0.724	0.500	0.038	0.178
500	Null	$1/n$	0.080	0.054	0.020	0.046
		$1/\log(n)$	0.074	0.060	0.046	0.042
		$1/\sqrt{\log(n)}$	0.078	0.040	0.024	0.038
	Alternative	$1/n$	0.874	0.824	0.028	0.276
		$1/\log(n)$	0.906	0.864	0.050	0.330
		$1/\sqrt{\log(n)}$	0.902	0.812	0.030	0.324
1000	Null	$1/n$	0.094	0.044	0.028	0.038
		$1/\log(n)$	0.086	0.032	0.020	0.030
		$1/\sqrt{\log(n)}$	0.064	0.042	0.032	0.040
	Alternative	$1/n$	0.958	0.968	0.022	0.576
		$1/\log(n)$	0.968	0.990	0.026	0.604
		$1/\sqrt{\log(n)}$	0.964	0.980	0.022	0.626

Table 7: Sensitivity analysis of Algorithm 1 for Scenario 3(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.160	0.070	0.026	0.070
		$1/\log(n)$	0.166	0.062	0.028	0.064
		$1/\sqrt{\log(n)}$	0.112	0.060	0.026	0.076
	Alternative	$1/n$	0.688	0.704	0.216	0.626
		$1/\log(n)$	0.688	0.726	0.204	0.636
		$1/\sqrt{\log(n)}$	0.698	0.722	0.210	0.644
500	Null	$1/n$	0.074	0.038	0.006	0.048
		$1/\log(n)$	0.076	0.064	0.030	0.038
		$1/\sqrt{\log(n)}$	0.084	0.054	0.030	0.038
	Alternative	$1/n$	0.722	0.792	0.250	0.692
		$1/\log(n)$	0.734	0.804	0.250	0.708
		$1/\sqrt{\log(n)}$	0.730	0.782	0.252	0.694
1000	Null	$1/n$	0.076	0.042	0.038	0.048
		$1/\log(n)$	0.064	0.034	0.034	0.020
		$1/\sqrt{\log(n)}$	0.060	0.038	0.032	0.036
	Alternative	$1/n$	0.766	0.822	0.216	0.738
		$1/\log(n)$	0.774	0.826	0.214	0.762
		$1/\sqrt{\log(n)}$	0.768	0.822	0.208	0.760

References

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

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.

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)* .

Mulzer, W. (2018). Five proofs of Chernoff's bound with applications. *arXiv preprint arXiv:1801.03365* .

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

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

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.

Zhu, J. & Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics* **14**, 185–205.

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*

$$E_P[\phi \mid N_1 = n_1, N_2 = n_2] \leq \alpha.$$

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*

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}_0} E_P[\phi] \leq \alpha \quad \textit{and} \quad \lim_{n\to\infty} \sup_{P \in \mathcal{P}_1} E_P[1 - \phi] = 0.$$

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*

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}_0} E_P[\tilde{\phi} \mid N_1 = n_1, N_2 = n_2] \leq \alpha \quad \textit{and} \quad \lim_{n\to\infty} \sup_{P \in \mathcal{P}_1} E_P[1 - \tilde{\phi} \mid N_1 = n_1, N_2 = n_2] = 0,$$

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

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

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.

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

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

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*

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

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.

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

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

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

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

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

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

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

(a) There exist constants $c, C > 0$ such that $\inf_{P\in\mathcal{P}_0} E_P(S_1^2) \ge c$ and $\sup_{P\in\mathcal{P}_0} E_P(S_1^{2+\delta}) \le C$ 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 $\|k\|_\infty \le K$.

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 Section 1.7. Putting things together concludes that T and $\tilde{T}$ are asymptotically equivalent.