Review Report

The manuscript provides several contributions to the problem of conditional two-sample testing. Firstly, it provides a theorem stating that conditional two-sample testing is a statistically hard problem in the sense that every valid conditional two-sample test (in the sense of type I error control for all null distributions) does not have power greater than its size against any alternative. This is a similar result to the hardness of conditional independence testing (Shah and Peters, 2020). By framing the conditional two-sample testing problem as a conditional independence testing problem, the hardness result essentially follows from the hardness of conditional independence testing, but this comes with some technical subtleties.

Secondly, the manuscript provides two general frameworks to construct conditional two-sample tests. The first one allows to turn an arbitrary conditional independence test into a conditional two-sample test and a theorem is provided that essentially states that level guarantees for the conditional independence test carry over to the conditional two sample test.

The second framework is based on density ratio estimation and allows to turn various techniques for marginal two-sample testing to conditional two-sample testing using importance weighting. The validity of this approach mainly hinges on the accuracy of the density ratio estimate and a classifier-based approach and an approach based on linear-time MMD are discussed in detail.

In my opinion, the contributions of the manuscript are interesting. Below, I provide some major comment regarding the presentation and various minor comments.

Presentation of the hardness result and the relationship to conditional independence tests

I found the treatment of the relationship between conditional two-sample tests and conditional independence tests somewhat confusing and arbitrary at some places. For example, the actual hardness result for conditional two-sample testing (Theorem 1) is already set up and formulated in the framework of a conditional independence test (in particular it includes the random variable Z taking values in $\{1, 2\}$). Only in the subsequent remark, it is argued that the type I error in the setting of the theorem is equivalent to the type I error when considering the conditional two-sample test. Would it not be more straightforward to formulate Theorem 1 really as a hardness result for conditional two-sample testing?

Also Theorem 2 considers the setting of conditional independence tests (with the random variable Z) but at the same time makes a statement about the output of Algorithm 1.

But Algorithm 1 is formulated in the setting of conditional two-sample testing (where there is no random variable Z a priori but two different datasets), so there is a mismatch and Algorithm 1 not well-defined in the setting of Theorem 2. Would it not be better to formulate Theorem 2 in terms of conditional two-sample tests, i.e. a statement about the validity of Algorithm 1 for conditional two-sample testing?

In my opinion, the distinction between the two settings should be made with more care. For me it helped to look at Theorem 2 and Algorithm 1 in the following way (I do not claim that this is the way it should be done, but I hope it brings across my point):

Setting 1 (cond. two-sample test): $n_1, n_2 \in \mathbb{N}$ fixed. Observe two mutually independent samples $\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \sim P_{XY}^{(1)}$ and $\{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \sim P_{XY}^{(2)}$. Sample random $\tilde{n}_1$ and $\tilde{n}_2$ as in Algorithm 1. If $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$, accept H_0 (output 0). Else, set $\mathcal{D}_{\tilde{n}} = \{(X_i^{(1)}, Y_i^{(1)}, 1)\}_{i=1}^{\tilde{n}_1} \cup \{(X_i^{(2)}, Y_i^{(2)}, 2)\}_{i=1}^{\tilde{n}_2}$ and run the conditional independence test ϕ (that has output either 0 or 1) using $\mathcal{D}_{\tilde{n}}$ at level α.

Setting 2 (cond. independence test): Let $n = n_1 + n_2$ and consider a random vector $(X, Y, Z) \in \mathbb{R}^{d_X + d_Y} \times \{1, 2\}$, where Z takes values 1 and 2 with probability n_1/n and n_2/n, respectively. Let $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$ be i.i.d. copies of (X, Y, Z) and let $N_1 = \sum_{i=1}^{n} 1(Z_i = 1)$ and $N_2 = \sum_{i=1}^{n} 1(Z_i = 2)$. Sample random $\tilde{n}_1$ and $\tilde{n}_2$ as in Algorithm 1. If $\tilde{n}_1 > N_1$ or $\tilde{n}_2 > N_2$, accept H_0. Else, let I_1 be the first $\tilde{n}_1$ indices i with $Z_i = 1$ and I_2 be the first $\tilde{n}_2$ indices i with $Z_i = 2$, set $\mathcal{D}_{\tilde{n}} = \{(X_i, Y_i, Z_i)\}_{i \in I_1 \cup I_2}$ and run the conditional independence test ϕ using $\mathcal{D}_{\tilde{n}}$ at level α.

Denote the output of setting 1 as ϕ_1 and the output of setting 2 as ϕ_2. Then, the distribution of ϕ_1 is equal to the distribution of $\phi_2 | N_1 = n_1, N_2 = n_2$. Note that setting 1 is the same setting as the problem setup in Section 1.1. and Algorithm 1. Setting 2 is the setup of Theorem 2 and what is meant with “the output of Algorithm 1” in Theorem 2. Maybe I am missing something, but I think that with the distinction into setting 1 and setting 2, one could formulate Theorem 2 and Algorithm 1 both in terms of setting 1 and then use the equality of the distributions of ϕ_1 and $\phi_2 | N_1 = n_1, N_2 = n_2$ for the proof.

Minor Comments

- The discussion of why one cannot in general apply a conditional independence test to the full data, but needs to reduce the sample size to $\tilde{n}$ (discussion between lines 266 and 285), seems important to me. Perhaps I am missing something, but the presented coupling argument is a bit unclear to me as it compares two datasets of the same size. How exactly is this connected to the question of why one cannot apply the conditional independence test to the full dataset? Why can’t one simply argue that since the number of discarded samples is $O((n\log(1/\epsilon))^{1/2})$, if the test is not changed too much by $O(n^{1/2})$ additional samples, it should be fine to use the full data? Also, Example 2 (unstable case) seems quite artificial. Is it possible to give another example or a better intuition? For example, from Section 3.5. in the Supplementary Material, one can see that RCIT (in contrast to GCM and PCM) is very sensitive to this issue. Is there an intuition why this is the case?

- Some typos / unclear formulations:
  - Section 1.4., line 191, undefined reference.
  - In Algorithm 1, shouldnt $\tilde{n} = k^{*}n$ be rounded to a whole number?
  - Line 272, I don't really understand what is meant with "meaning the conditional testing errors of ϕ are asymptotically equivalent to its marginal errors".
  - Line 302, shouldn't it be $\sum_{i=1}^{n} 1(Z_i = 1) = n_1$ instead of $\sum_{i=1}^{n} Z_i = n_1$ etc. in the definition of $\hat{f}$ and $\hat{g}$? (Z_i takes values in $\{1,2\}$ and not in $\{0,1\}$)
  - The expression on line 345 is always ≥ 0. How can it have expectation zero under the null hypothesis?
  - In equation (4), shouldn't 1 and 2 be swapped, since the true label for $V_i^{(1)}$ is 1 and the the true label for $V_i^{(2)}$ is 2?
  - I was first confused by the sampling mechanism in Section 5.2. Do I understand correctly that under the null hypothesis, the conditional distribution of $Y^{(1)}$ and $Y^{(2)}$) given $X^{(1)}$ and $X^{(2)}$, respectively is just the marginal distribution of Y? And for the alternative hypothesis, there is introduced some artificial dependence of the conditional distribution of $Y^{(2)}$ given $X^{(2)}$? If this is correct, it would perhaps be helpful to the reader to explicitly state that the dependence between X and Y in the original dataset is completely destroyed and some artificial dependence is introduced.