

Referee Report

This paper first establishes a hardness result and then introduces two frameworks for conditional two-sample testing. Below are several comments.

1. The hardness result is unsurprising, as it has already been discussed in Remark 4 of [Shah and Peters \(2020\)](#). Nevertheless, Theorem 1 and its proof require refinement. In the proof, instead of considering n i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} P$, the authors actually work with N i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^N \stackrel{i.i.d.}{\sim} P$. Then, under the event $\{N'_1 = \sum_{i=1}^N \mathbb{1}(Z_i = 1) \geq n_1, N'_2 = \sum_{i=1}^N \mathbb{1}(Z_i = 2) \geq n_2\}$, the authors use only n_1 observations with $Z = 1$ and n_2 observations with $Z = 2$ out of the N samples. Therefore, the resulting $n = n_1 + n_2$ data points are not i.i.d. from P , contrary to what is stated in Theorem 1. Moreover, the interpretation of the conditioning event $\{N_1 = \sum_{i=1}^n \mathbb{1}(Z_i = 1) = n_1, N_2 = \sum_{i=1}^n \mathbb{1}(Z_i = 2) = n_2\}$ in Theorem 1 is unclear, since the test ϕ is based on $n = n_1 + n_2$ data points out of N samples rather than n i.i.d. samples. In a nutshell, the test ϕ in the proof is a conditional independence test based on $\{(X_i, Y_i, Z_i)\}_{i=1}^N \stackrel{i.i.d.}{\sim} P$, which is not the same as that in Theorem 1.

Additionally, the proof crucially relies on Theorem 2 and Remark 4 of [Shah and Peters \(2020\)](#), where no assumption is made about the marginal distribution of $Z \in \{1, 2\}$, i.e., $P(Z = 1)$. However, the authors assume $P(Z = 1) = n_1/n$ for n_1, n_2 as given in Theorem 1, thereby imposing a restriction on the null distributions $\mathcal{P}_{0,M}$. This raises the question of whether the hardness result in [Shah and Peters \(2020\)](#) remains valid for conditional independence testing when the marginal distribution of $Z \in \{1, 2\}$ is specified. While the answer appears to be affirmative, at least when X is continuous, a discussion on the implications of this restriction would be valuable.

2. The statements in Theorem 2 are somewhat confusing. In Algorithm 1, the authors construct $\mathcal{D}_{\tilde{n}}$ of size \tilde{n} and then apply a conditional independence test ϕ using $\mathcal{D}_{\tilde{n}}$. However, in Theorem 2, the test ϕ is applied to $\{(X_i, Y_i, Z_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} P$, which consists of n samples rather than \tilde{n} . Besides, $\tilde{\phi}$ is a conditional two-sample test based on the given two samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$, rather than n i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} P$. As shown in the proof, the test $\tilde{\phi}$ can be expressed as $\tilde{\phi} = \mathbb{1}(\tilde{n}_1 \leq n_1) \mathbb{1}(\tilde{n}_2 \leq n_2) \phi_{\tilde{n}}$, where n_1 and n_2 are given. Consequently, the interpretation of the conditioning event $\{N_1 = \sum_{i=1}^n \mathbb{1}(Z_i = 1) = n_1, N_2 = \sum_{i=1}^n \mathbb{1}(Z_i = 2) = n_2\}$ with respect to $\tilde{\phi}$ is unclear. Further clarification is necessary to ensure correctness.
3. For the classifier-based approach, Assumption 1(b) is extremely restrictive. Even for a correctly specified parametric model \hat{r}_X (let alone more flexible nonparametric techniques), we have $E_P[\{\hat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2 | \hat{r}_X] = O_p(m^{-1})$, which implies that Assumption 1(b) is not satisfied. Consequently, unbalanced sample splitting becomes necessary, and taking $m = \lfloor n/2 \rfloor$ in practice is not justifiable. Notably, Assumption 2(b) in [Hu and Lei \(2024\)](#) is weaker than Assumption 1(b) in this paper; see the discussions after Theorem 1 in [Hu and Lei \(2024\)](#). A crucial question remains: how can

the sample be split in a data-adaptive manner when unbalanced sample splitting is unavoidable?

4. For the MMD-based approach, the authors adopt a linear-time MMD statistic instead of the quadratic-time MMD statistic. In marginal two-sample testing, the quadratic-time MMD test is known to achieve minimax optimal power. By contrast, the linear-time test generally exhibits lower power—by more than a constant factor—and has a worse rate (Ramdas et al., 2015), making it less favorable in practice. Although the standard permutation approach is not applicable to conditional two-sample testing, the authors should explore the strategies for calibrating the quadratic-time statistic, given its desirable theoretical properties.
5. No single test exhibits consistent performance across all scenarios, often displaying erratic size or power behavior in certain cases. A crucial question in practice, therefore, is: which test should one use to draw reliable conclusions? More specifically, which of the two frameworks should be preferred? If opting for the conditional independence-based framework, which conditional independence test should be employed? Alternatively, if selecting the density ratio-based framework, should one use the classifier-based test or the MMD-based test?
6. Two Gaussian distributions differing only in their means represent perhaps the easiest case, where a simple linear logistic regression model can correctly specify the density ratio. It would be valuable to assess the performance of the proposed tests when $X^{(1)}$ or $X^{(2)}$ is drawn from heavy-tailed or other non-Gaussian distributions.
7. In Section 5.2, since $X^{(2)}$ is sampled with probability proportional to $\exp(-x_1^2)$, the linear logistic regression (LLR) model misspecifies the density ratio. This misspecification may explain the poor performance of LLR on the superconductivity dataset. Nonetheless, an explanation of why LLR performs well on the diamonds dataset would be beneficial.
8. The medium heuristic is arguably the most popular method for selecting the bandwidth. Is there a specific reason why the bandwidth for the linear-time MMD test is fixed at 1 across all experiments?
9. On line 347, why does the expectation of the rank sum statistic equal zero under the null hypothesis?

References

- Hu, X. and Lei, J. (2024). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154.
- Ramdas, A., Reddi, S. J., Poczos, B., Singh, A., and Wasserman, L. (2015). Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.