General Frameworks for Conditional Two-Sample Testing
Referee Report

1 Summary

In this paper the authors consider the problem of conditional two-sample testing, where the goal is to test the equality of two conditional distributions $P^{(1)}_{Y|X}$ and $P^{(2)}_{Y|X}$, based on observed samples $\{(X^{(1)}_i, Y^{(1)}_i)\}_{i=1}^{n_1}$ and $\{(X^{(2)}_i, Y^{(2)}_i)\}_{i=1}^{n_2}$ from the corresponding joint distributions. The paper first establishes a hardness result for the conditional two-sample problem by relating it to conditional independence tests (CITs), which is known to be intrinsically hard. Then the authors propose two general frameworks:

- Converting existing conditional independence testing procedures to conditional two-sample test through a subsampling approach.
- Transforming the problem to a marginal two-sample testing problem between the joint distributions $P^{(1)}_{XY}$ and a weighted version of $P^{(2)}_{XY}$ where the weight is given by the density ratio between $X^{(1)}_1$ and $X^{(2)}_1$. To show applicability of this framework, the authors adapt the well known classifier based two-sample test and the kernel based two-sample test to this setting.

Furthermore, the authors justify their approaches with theoretical results and simulation studies.

2 Comments

The article tackles a significant and timely problem in statistical methodology with clear relevance to modern applications. The proposed frameworks, particularly their connections to conditional independence testing, are not only conceptually interesting but also open the door to leveraging powerful techniques from that area. Nevertheless, several aspects require clarification and justification.

2.1 Main Comments

- The core idea for using CIT based methods for the conditional two-sample problem is to use subsampling where on average $O\left(\sqrt{n\log(1/\varepsilon)}\right)$ many samples are discarded, which can affect the finite sample performance. The authors do acknowledge this limitation and provides experiments for choices of ε. I believe this requires a more thorough discussion, for example one can compare the power of the proposed test and the corresponding CIT in an oracle setting where we generate $Z \in \{1, 2\}$ and then generate $(X, Y)|Z \sim P^{(Z)}_{XY}$.
- The discussion about stability of the CIT method provided after Algorithm 1 needs more clarity. It would be beneficial for the reader if the authors clarify what they mean by stability when this is first mentioned. To my understanding it appears that the current discussion provides an alternative approach to Algorithm 1 where no samples are discarded.
- Building on the previous comment, this alternative approach seems underdeveloped. While Algorithm 1 and Theorem 2 provides a general CIT based framework, this alternative approach seems highly dependent on the specific CIT method and its implementation. A case-

by-case evaluation seems insufficient for building a general framework and it would be great to have more concrete guidelines for this alternative approach.

- Additionally, Example 1 assume knowledge of conditional expectations and Example 2 is an extreme case instability under estimation. I think a more detailed discussion is required on how estimation affects instability and as a consequence the alternative approach.

- Theorem 1 establishes that conditional two-sample testing is generally impossible (achieving non-trivial power) without further assumptions, implying that structural constraints (for example, smoothness conditions) are necessary. In the density ratio based approaches (henceforth abbreviated as DRT), it is not immediately clear how such structural assumptions are implicitly incorporated. It would be great if the authors can discuss how the assumptions for DRT (Assumptions 1 and 2, in particular to my understanding 1(b) and 2(b)) serve as necessary structural conditions to overcome the hardness from Theorem 1.

- Under the DRT setting, Assumptions 1(b) and 2(b) place restrictions on the convergence rates of the density ratio estimator. It would be beneficial for the reader if there is a discussion about whether these rates are achievable in practice, citing relevant sources and the corresponding estimators which achieve these rates.

- One of the specific DRT approaches is based on the kernel based MMD statistic. However, the authors only discuss the linear MMD statistic. It is well known that the quadratic MMD achieves much better power than the linear one. While the authors acknowledge the difficulty of implementing the quadratic one using permutation tests, is it possible to have a version of the quadratic MMD test and calibrate that using multiplier bootstrap as in Gretton et al. [2009]?

- Building on the previous point, the manuscript's discussion of permutation-based implementations for the quadratic-time MMD feels incomplete. Several permutation strategies already exist for conditional independence tests—for example, Berrett et al. [2020], Candes et al. [2018], Kim et al. [2022]. Because Theorems 1 and 2 establish a link between conditional independence tests and conditional two-sample tests, it seems natural to ask whether those existing strategies could be adapted to produce a permutation test in the quadratic-time setting. If such an adaptation is difficult, a detailed explanation of the conceptual or computational obstacles would be helpful.

- Building on the authors' classifier-based conditional two-sample tests, recent work has focused on *local* conditional comparisons—specifically, testing whether $P^{(1)}_{Y|X}$ equals $P^{(2)}_{Y|X}$ at a given covariate value $X = x_0$ (see Linhart et al. [2023]). Could the proposed framework be adapted to develop such point-wise conditional tests?

2.2 Minor Comments

- There appears to be some minor inconsistency between Theorem 2 and Algorithm 1. Specifically, Algorithm 1 defines Z_i for $1 \leq i \leq \tilde{n}_1$ and similarly for $\tilde{n}_2$, but Theorem 2 defined N_1 and N_2 using the full sample size n. I am not sure what happens for the discarded samples as those Z_i are never defined.

- At first glance, Theorem 2 is framed as an asymptotic result. Yet any finite-sample valid CIT ϕ would appear to satisfy the same assumptions. If that is correct, could the guarantees for Algorithm 1 be reformulated to hold in finite samples as well? A brief remark following Theorem 2 would help clarify this point.

- For completeness, please include the explicit refined bound for $k^\star$ in the supplementary material.
- A concise discussion of the computational costs associated with the different methods would be very helpful for practitioners.
- A reference link seems to be broken at the end of page 5.
- Minor proofreading for typos/grammar is recommended.

References

T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197, 2020.

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, 22, 2009.

I. Kim, M. Neykov, S. Balakrishnan, and L. Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022.

J. Linhart, A. Gramfort, and P. Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference. *Advances in Neural Information Processing Systems*, 36:56384–56410, 2023.