Response to reviews of "On the power of conditional independence testing under model-X"

EJS2107-037

Eugene Katsevich and Aaditya Ramdas

October 7, 2022

1 Overview of Revisions

1. We found that Proposition 1 was incorrect. Therefore, Theorem 1 now states optimality among conditionally valid tests, rather than among all (marginally) valid tests.

2. Added references to recent methods papers (Asher, Grunwald) inspired by our results.

2 Responses to Referee 2

I find the revision of the paper very much improved, and thank the authors for carefully addressing the comments made in the previous round. I only have some minor comments listed below.

2.1 Minor points / suggestions

1. *It could be helpful to be a bit more precise about the set of distributions being considered in (9). The notation and proofs suggest you are only considering distributions that are absolutely continuous with respect to Lebesgue measure (or you are imagining some dominating measure). It would be helpful to make this clearer.*

   Indeed, we had been implicitly assuming that all distributions have densities (and conditional densities) with respect to a dominating measure, and identifying distributions with their densities in our notation. The earliest point in the paper where we do this is actually equation (3). We have added the following footnote to this equation to make our assumptions explicit:

> We implicitly assume that $\mathcal{L}$ has a density with respect to some dominating measure on $\mathbb{R}^{1+1+p}$, and that all conditional densities are well-defined almost surely. Here and throughout the paper, we identify probability distributions with their densities with respect to the appropriate dominating measure.

We acknowledge that this statement is still rather informal, but prefer to avoid the measure-theoretic technicalities involved in defining conditional densities.

2. *It is perhaps also worth being a bit more formal in Proposition 1 where I believe that strictly speaking the supremum over y, z should be removed and the inequality should only hold almost everywhere. This point applies analogously to all the theoretical results.*

   As discussed in the overview of the revisions (Section 1 above), we have removed Proposition 1 from the manuscript. Nevertheless, we agree with the point that the suprema over y and z should be replaced by almost sure statements. This is reflected in equation (11) in the main text and in a few modifications to the proof of Theorem 1 in Appendix A.

3. *It is perhaps worth mentioning that (18) is a special case of a partially linear model.*

   We have done so parenthetically above this equation.

4. *In the statement of Thm 2, the α subscript of C is replaced with "n" which is perhaps a bit confusing. Perhaps this could be commented on, or alternative would be to carry the α notation in some way.*

   We have dropped the α from C_α, noting that the dependence on α is implicit.

5. *I would suggest changing the g_n' notation as it may be mistaken for a derivative.*

   We have replaced g_n' by $\bar{g}_n$ to avoid this confusion.

6. *The moment conditions in Thm 4 seem unnecessarily strong. I don't think it is critical to weaken them, but for instance I think one does not need a second moment condition for a triangular array weak law of large numbers to hold. 1+δ absolute moments should suffice (e.g. lemma 19 of https://arxiv.org/pdf/1804.07203.pdf). Similarly the application of the Lyapunov CLT can use an abitrary $\delta > 0$ rather than $\delta = 1$. Perhaps these sorts of observations can slightly weaken the moment conditions.*

   We agree with this observation, and we have added a footnote to page 12 stating that

   > The exponents in these moment conditions can be relaxed from 4 to $2 + \delta$.

At this stage, we have not updated the actual statements and proofs of our theoretical results to reflect this change, in part because we could not find a reference (either textbook or paper) stating the triangular array weak law of large numbers with $1 + \delta$ moments. In particular, note that the result of Shah and Peters cited by the referee is not in the triangular array setup. Therefore, we feel the slight relaxation of our moment assumptions may not be worth the extra effort required. If the referees disagree, we would be willing to add a statement and proof of the aforementioned weak law of large numbers and update our theorem statements and proofs accordingly.

7. *pg 24 I feel the sentence starting "Knockoff inference is then based on a form of data-carving..." is a bit too long and could be broken into two.*

   We have made this change.

8. *Thm 5: Is the optimal one-bit p-value essentially unique? if not, then perhaps it is worth modifying the discussion a little here and elsewhere to not refer to *the* optimal test*

   We are not sure about the uniqueness of the optimal one-bit p-value, so we have changed the language to avoid referring to "the optimal test."

3 Responses to Referee 3

3.1 Summary

This manuscript studies some theoretical properties of model-X inference and of approximate versions of the conditional randomization test under certain semi-parametric assumptions. The main two issues explored here are: (1) the optimal choice of test statistics for CRT and knockoffs; (2) the interface between the CRT and a modified version of the GCM test of Shah & Peters.

3.2 Overall impression

This manuscript contains interesting ideas and it is clear that a good amount of work went into its preparation. I agree with the authors that this revision involves some improvements compared to the first submission, and I am glad to see the comments provided in the first round of review have been taken into (some) account. However, some concerns raised in the first round of review are still relevant. The paper still feels a bit disconnected and somewhat incomplete. I think I learned something by reading it, and yet I feel the abstract over-promised and in the end I have more questions than answers. These issues may not necessarily be fatal, but I would take them into consideration when deciding whether the paper could benefit from further revision.

3.3 Major comments

There are two interesting methodological/theoretical themes in this paper: (1) the optimal choice of test statistics for model-X testing, and (2) the robustness of model-X testing to second-order approximations. These two ideas could be developed further into two very nice separate papers, but they don't fit very well together now and they have not yet been explored to fully satisfactory depth.

This is a very similar concern to what the referee expressed last time. We did a decent bit of rewriting last time to de-emphasize the robustness part, and to emphasize the unifying theme of "power of model-X conditional independence tests." Perhaps the remaining issue is that Section 3 is still a bit of a detour from this main theme, despite our efforts in the first paragraph of that section to frame it as a prelude to the following section on CRT power. Navigating this issue may require some thought.

Regarding theme (1), Section 2.1 is interesting but very unsurprising. I would almost say Proposition 1 was not stated explicitly in prior literature because it is obvious. The authors seem to admit so in the last paragraph of Section 2.1, but perhaps this could have been pointed out sooner.

Sure, we can point it out sooner.

My only technical concern here is that there may be a typo, or perhaps just some ambiguity, in the notation of its proof. I was expecting to see a delta function in the integrand. It could be a typo, or it could be that I misunderstood something in the notation of the proof (I'm quite sure the result is correct). In any case, Section 1 is valuable as a prelude to Section 2.2, which contains the less obvious Theorem 1.

I believe the proof of Proposition 1 is correct as stated.

Theorem 1 is also very intuitive, as it follows quite directly from the Neyman-Pearson lemma, but it is definitely worth stating explicitly. My main concern with theme (1) is that this line of inquiry seems to terminate abruptly and prematurely after the statement of Corollary 1. What should we do with this Neyman-Pearson result? Does it have any practical relevance (as we don't deal with point alternatives in practice)? Does it really answer any questions about which statistics should be used in practice? Does it provide "quantitative explanations for empirically observed phenomena and novel insights to guide the design of MX methodology", as we were promised in the abstract?

We address the question of the practical relevance of Theorem 1 / Corollary 1 in the paragraph following Corollary 1. Admittedly, the referee is right in stating that this result, on its own, does not directly inform application of MX methodology, though we point to one or two possibilities for practical insights. In retrospect, I would formulate this result as a basic optimality result for MX testing, rather than a guide for practical application. I would agree with the referee that "novel insights to guide the design of MX methodology" is probably an oversell.

Regarding theme (2), the second part of the paper focuses on the partially linear model, but there is no mention of that in the abstract. A partially linear semi-parametric model within the CRT framework is almost as big of an assumption as to assume a fully linear model, because the Z variables are essentially conditioned upon (Section 2.1). Then, if we believe in a linear model, do we still need model-X inference?

The referee has misunderstood: We are using the partially linear model only as

an alternative distribution against which we assess power, rather than an assumption required for validity of the MX(2) F-test. I thought we had included wording to clarify this nuance, but I can't find it in the manuscript.

In any case, what is the take-away message? Is there enough evidence to recommend practitioners to broadly utilize the MX(2) F-test (or the closely related GCM test) instead of the model-X test? The experimental section of this paper feels a bit shallow; clearly, this is a more theoretical paper. Yet, the theoretical analyses involve asymptotic approximations and a partially linear model which seem to go against the very core principles of model-X testing (finite-sample guarantees and no assumptions on Y—X). Theoretically, this feels counter-intuitive. For example, in the introduction the authors discuss GWAS as a success story for model-X testing. Do the results in this paper specifically explain why model-X testing has so far shown promise in GWAS? Do the results in this paper suggest instead that the MX(2) F-test can be safely applied to GWAS? What are some of the possible limitations or concerns of the MX(2) F-test in the applications that model-X testing has so far primarily focused on?

We stated in our previous rebuttal that "we view our contribution as primarily theoretical, establishing basic power and optimality results for increasingly popular methodologies like the CRT." I think we need to continue stressing this point. We should push back on the statement that "the theoretical analyses involve...a partially linear model which seems to go against the very core principles of model-X testing"; see the red text above. As far as actual deployment of the MX(2) F-test as a *methodology* (e.g. for GWAS analysis) as opposed to as a *theoretical device*, I would reserve judgment or speculation for a future work. This is the reason why our numerical results section is not as fleshed out as in typical methods papers.

3.4 Minor comments

- *Proof of Theorem 1: In equation (61), have P_0 and P_1 been defined before?*

  This notation was a bit sloppy on my part. I can clarify this easily.

- *On page 12, there is a typo "trained independent" below equation (31)*

  Will fix.

- *Theorem 2: should the partially linear model assumption be stated explicitly (if needed)?*

  Again, the referee incorrectly understood that we are actually using the partially linear model assumption.

- *Bottom of page 12: the conjecture seems quite strong. If some evidence will be provided later, it could be anticipated here.*

  FYI, A version of this conjecture will be proved in a paper I have coming out soon. Perhaps I can just add another (handwavy) sentence providing intuition.

- *Algorithm 2: Line 1 seems a bit vague (a mix of both). Should g-hat be fitted in sample or out of sample? This is a big recurring question in this paper but is not answered.*

  The paper I have coming out soon shows that, if g-hat is sufficiently accurate, it's ok to fit it in sample in the context of the dCRT (even if there's some error in $\mathbb{E}[X|Z]$), like the GCM test. For this paper, I think we have to just leave this as a question for future work.

- *Last paragraph of Section 3.3. Again, this conjecture seems quite strong. Is there enough evidence to support it? Does it really never matter at all how g-hat is fitted?*

  Again, I think we have to largely punt to future work. We could comment that under the MX(2) assumption, the quality of $\widehat{g}$ impacts power but not Type-I error control.

- *Section 3.4. My impression is that more effort could have gone into the simulations. Why try to validate empirically the robustness of MX(2) F-test but not do any experiments about anything else? In any case, I wouldn't say MX(2) F-test has been validated super-extensively.*

  I think the confusing part is that we are sometimes treating the MX(2) F-test as a *methodology* (e.g. when we give it a name and an algorithm box) and other times as merely a *theoretical device*. When viewed as a methodology, our numerical simulation does seem insufficient. When viewed as a theoretical device, you could say our numerical simulation is perhaps even more than people might typically do. We're walking a bit of a fine line here and trying to have it both ways, which is what the referee is picking up on.

- *How does MX(2) F-test compare to the GCM test in practice?*

  Similar to the above comment: we are not really trying to propose the MX(2) F-test as a methodological alternative to the GCM. It's purpose is mainly as a theoretical device. But then again, why give it a name and an algorithm box?