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

EJS2107-037

Eugene Katsevich and Aaditya Ramdas

October 22, 2022

1 Overview of Revisions

We thank the referees for their constructive comments, which we have incorporated into a further improved manuscript. The revisions includes the following updates:

Updated Theorem 1. Proposition 1 in the previous version of the manuscript stated that any valid MX CI test must also be conditionally valid. Unfortunately, we discovered a mistake in our proof of this fact, and subsequently found counterexamples. Therefore, we have removed Proposition 1 from the paper and formulated Theorem 1 as a statement about the most powerful conditionally valid test (and in particular the most powerful CRT) rather than the most powerful MX CI test overall. We leave it as an open question whether the most powerful MX CI test is in fact a CRT.

Added references to recent papers inspired by our results. Recently, Grunwald et al (2022) proposed an extension of the CRT to the sequential setting. They proposed a test relying on a likelihood-based statistic, citing inspiration from our Theorem 1. Similarly, Spector and Fithian (2022) proposed a more powerful version of knockoffs based on a *masked likelihood ratio* statistic, citing inspiration from our Theorem 5. We have added references to these two papers to our revised manuscript to demonstrate how our work has inspired new model-X methodologies.

Moved the numerical simulation section to the appendix. Over the course of the last two rounds of revisions, we have focused the paper on the subject of power of MX CI tests, and framed the MX(2) F-test and associated questions of robustness as secondary contributions. In line with this, we have moved the numerical simulations of the MX(2) F-test to Appendix B. This way, readers can move more quickly from Section 2 (power of the CRT in finite samples) to Section 4 (power of the CRT in large samples).

These and other updates are highlighted in blue in the revised manuscript for the reviewers' convenience. Below, we address the referees' comments point by point.

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.

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.

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.

We are happy to have improved the manuscript based on the referee's helpful comments.

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.

We hope that, in the latest revision, we have addressed the referee's remaining concerns. Please see below.

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.

We understand how these two themes appeared disconnected in initial versions of the manuscript. In our responses to the referee's comments in the previous round, we stated that

> "We thank the referee for pointing out the somewhat disconnected nature of the submitted manuscript, especially the transition from Section 2 to Section 3. We agree, and we have substantially edited the exposition in Section 3 for a smoother flow...Ultimately, the central goal of the manuscript is to study the power and optimality of the CRT, and the MX(2) F-test is mostly a means to that end—though we argue it is of independent interest as well. This logic is conveyed in the opening paragraph of Section 3 (p. 9)..."

We agree that Section 3 still seems like a detour from our primary focus of studying the power of the CRT. To this end, we have moved the former Section 3.4 (the numerical simulation for the MX(2) F-test) to Appendix B. The MX(2) F-test and the question of robustness are not our primary focus, so we hope relegating the simulation to the appendix makes the main text more focused and streamlined. Furthermore, we have updated the last paragraph of Section 3 for a smoother transition to Section 4:

"Having found that the dCRT is a natural test to apply for power against semiparametric alternatives, and that this test is equivalent to the simpler MX(2) F-test, we are ready to study the relationship between the power of the dCRT and the quality of the underlying machine learning procedure."

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. 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.

We appreciate the referee's suggestion for us to check the proof of Proposition 1. As it turns out, the proof was in fact wrong (though for a different reason than the referee suggested). Furthermore, we found the proposition itself to be false; please see the overview of the revisions (Section 1) for more on this. What can be stated instead of Proposition 1 is that every CRT is a conditionally valid test (as opposed to every marginally valid MX CI test is also conditionally valid). We state this in Section 2.1 of the revision, but because it is obvious we do not even formalize this statement as a proposition. We mention in the second sentence of Section 2 that the observation of conditional validity of the CRT was implicit in earlier works, and later in the section that the existence of a single null distribution from which to resample $\widetilde{X}$ is central to the very definition of the CRT. We view Section 2.1 as primarily setting the stage for the optimality result in Section 2.2, rather than a presentation of new results.

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?

Our primary theme of power of MX CI tests is addressed not only in Section 2 (finite-sample power of CRT), but also in Section 4 (asymptotic power of CRT) and Section 5 (finite-sample power of knockoffs). As far as the consequences of the Neyman-Pearson result in Section 2, we discuss in the paragraph following Corollary 1 that this result provides evidence in favor of likelihood- or loss-based test statistics, and discuss other advantages of using such statistics. At least two recent papers have taken methodological inspiration from our optimality results (Grunwald et al, 2022 for CRT and Spector and Fithian, 2022 for knockoffs). We have acknowledged these works in this revision in

Section 2:

> "Theorem 1 has inspired an extension of the CRT to the sequential setting using a likelihood-based test statistic, accompanied by a similar optimality result (Grunwald et al, 2022)."

and Section 5:

> "Recently, Theorem 5 inspired a more powerful variant of knockoffs based on *masked likelihood ratio* statistics, which comes with a different kind of optimality guarantee (Spector and Fithian, 2022)."

Finally, we have updated the sentence in the abstract referenced by the referee to make it more concrete:

> "In this paper, we study the power of MX CI tests, yielding quantitative insights into the role of machine learning and providing evidence in favor of using likelihood-based statistics in practice."

Here, "quantitative insights into the role of machine learning" refers to Theorem 4 (connecting the performane of the machine learning method to the power of the CRT) and "evidence in favor of using likelihood-based statistics in practice" refers to Theorem 1 (showing that the likelihood test statistic is finite-sample optimal).

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?

It seems the referee has understood the partially linear model to be an additional assumption we make for the validity of inference. We have added a sentence immediately after the introduction of this model to clarify that this is not the case:

> "We emphasize that—in this section and throughout the paper—we use the partially linear model (17) exclusively as an alternative distribution against which to assess power, rather than an additional assumption required for Type-I error control."

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 fully agree that the MX(2) F-test is not fully fleshed out or tested as a methodology. As we suggest in a few places in the manuscript, for our purposes the MX(2) F-test is primarily a stepping stone to the analysis of the dCRT, rather than a methodology we advocate for immediate practical use. For example, in the introduction, we position our results on the MX(2) F-test as auxiliary:

> "On the way to addressing Q2, we additionally establish a third result (Section 3) that may be of independent interest: **A resampling-free second-order approximation to the dCRT is equivalent to the dCRT and controls Type-I error under weaker assumptions.**"

Our choice to move the MX(2) F-test simulations to the appendix in this revision also reflects this. A fuller methodological treatment of the MX(2) F-test is beyond the scope of the current paper, and accordingly we reserve judgment on its practical utility for applications. Finally, we point out that the validity of the MX(2) F-test *does not* rely on a partially linear model assumption; as stated above, we use partially linear models only as alternatives against which to evaluate power.

3.4 Minor comments

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

  Indeed, the notation had not been previously defined. We have updated the equation to remove this notation altogether, as the proof is still clear without it.

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

  Thank you for pointing out this typo; we have fixed it.

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

  See the discussion above: we use the partially linear model exclusively as an alternative distribution against which to assess power, rather than an additional assumption required for Type-I error control.

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

  We have added a sentence after the conjecture with our intuition for why it might hold:

  > "At least, we observe that the conditioning in the construction of the resampling distribution $\mathcal{L}_n(\sqrt{n} \cdot \widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z)$ ensures that its

mean and variance remain equal to 0 and $\widehat{S}_n^2$ even when $\widehat{g}_n$ is fit in sample."

- *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.*

  While we conjecture that the MX(2) F-test continues to be valid for $\widehat{g}_n$ fit in sample, we acknowledge that at this stage this is just a conjecture. Therefore, in the algorithm box we have specified that $\widehat{g}_n$ is fit out of sample. It is unfortunately beyond the scope of the current work to theoretically investigate the properties of the MX(2) F-test in the case when $\widehat{g}_n$ is fit in sample.

- *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?*

  In addition to the numerical simulation supporting this conjecture, we have added a sentence of intuition:

  > "One may expect this because the validity of the MX(2) F-test derives from the correctness of (μ_n, Σ_n) rather than that of $\widehat{g}_n$."

  Indeed, the spirit of model-X inference is to put the modeling burden on $\mathcal{L}(X|Z)$ rather than on $\mathcal{L}(Y|Z)$. Given that $\widehat{g}_n$ is a property of the latter, it makes sense that we can maintain Type-I error control without too many requirements on $\widehat{g}_n$.

- *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.*

  We agree that our validation of the MX(2) F-test falls far short of the threshold required to advocate for its practical use. As we discuss above, for the sake of this paper the MX(2) F-test is primarily an auxiliary construction used to help understand the dCRT, though it may be of independent interest. Investigating its practical performance is beyond the scope of the current paper.

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

  Please see the answer to the previous question. We defer the comparison of the MX(2) F-test to alternatives like the GCM test to future work.