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

EJS2107-037

Eugene Katsevich and Aaditya Ramdas

March 2, 2022

We would like to thank the referee and the associate editor for useful comments and for the opportunity to revise our paper. Based on these suggestions, we made several improvements to the revised manuscript. We summarize these next, and then provide point-by-point responses to the comments.

1 Overview of Revisions

To sum up the referee's main concerns, (a) the manuscripts's aim are somewhat disconnected from each other and (b) the robustness and methodological portions of the manuscript are weaker than the other portions. We agree with these assessments, and have made revisions to address both of them. Most importantly, **we restructured the paper around the main goal of studying fundamental power and optimality properties of model-X methods.** This simultaneously addresses the two main concerns by (a) by making the paper's exposition cleaner and more focused and (b) placing more emphasis on our primary contributions while placing less emphasis on our secondary contributions, the latter including the robustness and methodological aspects. In particular:

- We updated Sections 1.2 and 1.3 in the introduction to reflect the focus on power and optimality of MX methods.
- We rewrote Section 3 to emphasize that the primary goal of the MX(2) F-test is a stepping stone to the power analysis of the CRT, while discussing the methodological and robustness implications as secondary consequences of independent interest.

Additionally, we bolstered the methodological aspect of the paper as follows:

- We clarified the relationship between the MX(2) F-test, the dCRT, and the GCM test in Section 3.5, including Tables 1 and 2. We discussed that the MX(2) F-test may be preferred to the dCRT due to its speed and that the assumptions necessary for the MX(2) F-test and GCM test to control Type-I error do not subsume each other. We argue that all three tests have their relative strengths and weaknesses.

- We conjectured that the MX(2) F-test continues to control the Type-I error even if the machine learning step is done in sample (Section 3.3), and updated the numerical simulation to provide support for this conjecture (Section 3.4). If true, this conjecture would make the MX(2) F-test more directly competitive with the dCRT and GCM tests.

Finally, we added several smaller revisions (primarily in Section 3) to fully address all the referee's comments. We respond point by point to these comments next.

2 Responses to Referee

2.1 Major comments

1. *I do find the different sections and aims of the paper a bit disconnected— specifically I'm not sure if (1) is well-connected with (2) and (3). (1) is to do with the power of the CRT test. On the other hand, (2) is about a different test that works with different assumptions, though there is an asymptotically equivalent CRT test, and (3) studies the power of this test. The motivation behind (2) seems to be concerns about the validity of the CRT assumptions (which require full knowledge of the conditional distribution $X|Z$), but I don't see how this idea of 'robustness' really relates closely to (1). One option could be to remove or de-emphasize (1)—in my opinion it is nice to have such a result written down somewhere, but it is unsurprising.*

   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. The high-level logic in the revised manuscript is that Section 2 studies the power and optimality of the CRT in finite samples against point alternatives while Sections 3 and 4 study the power of the CRT in an asymptotic setting against semiparametric alternatives. Section 3 prepares the groundwork by establishing the asymptotic equivalence between the distilled CRT and the simpler-to-analyze MX(2) F-test (loosening the assumption from MX to MX(2) along the way), while Section 4 establishes the power of the MX(2) F-test and therefore the distilled CRT. 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 (pp. 8-9):

> "In Section 2, we saw how to construct the optimal test against point alternatives specified by $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. In practice, of course we do not have access to this distribution, so we usually estimate it via a statistical machine learning procedure. The goal of this section and the next is to quantitatively assess the power of the CRT as a function of the prediction error of this ML procedure. Specifically, we consider the power of a specific instance of the CRT against a set of semiparametric alternatives (Section 3.1). We prepare to assess the power of this test by showing its asymptotic equivalence to the simpler-to-analyze *MX(2) F-test* (Section 3.2), which is of independent interest due to its closed form and weaker assumptions (Section 3.3). We examine the finite-sample Type-I error control of the MX(2) F-test in numerical simulations (Section 3.4) and put this section's results into perspective (Section 3.5) before moving on to stating the desired power results in the next section (Section 4)."

2. *Regarding the robustness of the CRT test, I feel one could argue that assuming exact knowledge of the first two conditional moments of $X|Z$ is not too much weaker than asking for knowledge of the entire distribution. Are there specific examples where completely correct models for the conditional means and variances are known, but not the conditional distribution? In my opinion, a more relevant analysis would quantify the impact of the error in estimating these conditional moments. Also, it might be interesting to consider a version that does not require estimating the second conditional moment, but potentially places a slightly stronger condition on the error in estimating $\mathbb{E}[Y|Z]$.*

   We agree with the referee that assuming exact knowledge of the first two moments of $X|Z$ is still a strong assumption, which we would not necessarily expect to be satisfied in practice. We nevertheless find it comforting that the MX methodologies considered here are not sensitive to misspecification of the higher moments of the assumed $X|Z$. We agree that a stronger kind of robustness analysis would quantify the impact of the error in estimating these conditional moments and/or shift more of the assumptions onto $\mathbb{E}[Y|Z]$, but these directions are beyond the scope of the current work. We convey these points in Section 3.5 of the revised manuscript:

   > "...Theorem 3 shows that asymptotic Type-I error control of MX-style methodologies can be achieved under the weaker MX(2) assumption, requiring only two moments of the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ rather than the entire conditional distribution...note that the MX(2) assumption still requires *exact* knowledge of the first and second conditional moments; we leave as an important future direction to examine the robustness of these tests to errors in these quantities. Note that bounds on the worst-case Type-I error of the CRT have been obtained in the context of a misspecified $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ (Berrett et al 2020)."

3. *What is the significance of the asymptotic equivalence of the CRT test constructed using the test statistic on which the MX(2) F-test is based, and the MX(2) F-test*

itself? I cannot see why the CRT test would be preferred here? Overall, the connection between CRT tests and the MX(2) F-test seems quite weak to me.

The significance is that we can study the power of the CRT by instead studying the power of the easier-to-analyze MX(2) F-test. We have clarified this motivation to construct the MX(2) F-test in the first paragraph of Section 3 (quoted above). Furthermore, we expand on the methodological advantages and disadvantages of the MX(2) F-test compared to the CRT in Section 3.5 of the revised manuscript:

> "The preceding results suggest that the MX(2) F-test is a useful alternative to the dCRT: the power of these methods is asymptotically the same (Theorem 2), while the MX(2) F-test is computationally faster because it does not require resampling (Table 1). On the other hand, note that we have proven Type-I error control for the MX(2) F-test only when $\widehat{g}_n$ is fit out of sample and only asymptotically, while the dCRT gives finite-sample Type-I error with in-sample fit $\widehat{g}_n$. However, numerical simulations suggest good finite-sample Type-I error control for the MX(2) F-test even when $\widehat{g}_n$ is fit in sample."

4. *The methodological contribution here is quite modest as the proposed MX(2) F-test is essentially a multivariate version of the Generalized covariance measure proposed by Shah & Peters.*

We view our contribution as primarily theoretical, establishing basic power and optimality results for increasingly popular methodologies like the CRT. Furthermore, we agree that the MX(2) F-test has a close relationship to the GCM test. Nevertheless, the MX(2) F-test complements the GCM test in the sense that it controls Type-I error in regimes when the latter is not guaranteed to, e.g. in cases when $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ cannot be consistently estimated. We compare the two methodologies in Section 3.5 of the revised manuscript and in Tables 1 and 2, reproduced here:

> "To compare our results with those in non-parametric doubly robust inference, we consider the closest representative of the latter: the generalized covariance measure (GCM) test of Shah and Peters. The GCM test statistic is defined as
>
> $$\widehat{\rho}_n^{\mathrm{GCM}} \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \widehat{\mu}_n(Z_i)), \tag{1}$$
>
> where $\widehat{\mu}_n(\boldsymbol{Z})$ and $\widehat{g}_n(\boldsymbol{Z})$ are estimates of $\mu_n(\boldsymbol{Z}) = \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $g_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$, respectively. This statistic is shown to converge under conditional independence to a mean-zero normal limit as long as the estimates of $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$ are both consistent, while the product in these estimation errors tends to zero at a rate of $o(n^{-1/2})$. By contrast,

the MX(2) F-test places more weight on the model for $\boldsymbol{X}|\boldsymbol{Z}$ (assuming both first and second moments of this conditional distribution are known) while placing less weight on the model for $\boldsymbol{Y}|\boldsymbol{Z}$ (not assuming even consistency for $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$). Therefore, while the MX(2) F-test closely resembles the GCM test, the assumptions required for validity of these two methods do not subsume each other (Table 2)."

Method	Guarantee	Resampling
CRT	Finite-sample	Yes
MX(2) F-test	Asymptotic	No
GCM test	Asymptotic	No

Table 1: Type-I error guarantee and necessity of resampling for each method compared.

Method	$\mathcal{E}(\mathbb{E}[\boldsymbol{X}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathrm{Var}[\boldsymbol{X}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathcal{L}(\boldsymbol{X}\|\boldsymbol{Z}))$	$\mathcal{E}(\mathbb{E}[\boldsymbol{Y}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathbb{E}[\boldsymbol{X}\|\boldsymbol{Z}]) \times \mathcal{E}(\mathbb{E}[\boldsymbol{Y}\|\boldsymbol{Z}])$
CRT	0	0	0	–	–
MX(2) F-test	0	0	–	–	–
GCM test	$o(1)$	–	–	$o(1)$	$o(n^{-1/2})$

Table 2: Assumptions necessary for each method compared (excluding moment assumptions). Here, $\mathcal{E}(\cdot)$ refers to the root-mean-squared estimation error of a given quantity.

2.2 Other comments

1. *In equation (48), the RHS doesn't depend on j.*

   While this is not a typo, the result does look strange. Intuitively, the optimal test of variable j should of course depend somehow on j. However, as we point out in the revised manuscript, the *test* (i.e. the one-bit p-value) does in fact depend on j, even if the *test statistic* does not (see excerpt from p. 25 of the revision below). This is an artifact of the definition of the one-bit p-value of knockoffs.

   > "The reader observes that the optimal test statistic is not a function of the knockoff variables or of the index j, which may at first seem paradoxical. Recall from the definition (53), however, that the one-bit p-value compares the test statistic on the original augmented design $[X, \widetilde{X}]$ and its swapped version $[X, \widetilde{X}]_{\mathrm{swap}(j)}$. Therefore, the optimal one-bit p-value checks whether the original jth variable $X_{\bullet,j}$ fits with the rest of the data better than does its knockoff $\widetilde{X}_{\bullet,j}$. Therefore, the optimal one-bit p-value is in fact a function of the knockoffs as well as the index j."

2. *I was surprised that the non-bold Z is a matrix, while the bold Z is a vector. I would have expected it to be the other way round.*

There are a variety of conventions in use for denoting random variables versus their realizations as well as scalars versus vectors versus matrices, and we acknowledge that our notation is inconsistent with some such conventions. However, we found that the most important notational distinction to make in our paper was between population-level random quantities and their sample-level realizations; we denoted these by bold and non-bold symbols, respectively. Our notation does not distinguish between scalars, vectors, and matrices, as the presence of subscripts disambiguates the object type. For example, $Z \in \mathbb{R}^{n \times p}$ is the entire matrix of covariates whereas $Z_i \in \mathbb{R}^{p \times 1}$ is one row of this matrix. We have added the blue text below to the notation section on p. 5 of the revision to clarify this:

> "Recalling equations (1) and (2), population-level variables (such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$) are denoted in boldface, while samples of these variables (such as X_i, Y_i, Z_i) are denoted in regular font. Note that boldface does *not* distinguish between scalars, vectors, and matrices, as it is sometimes employed."

3. *Where is the $X_{\bullet,j}$ etc. notation defined that is used in Section 5?*

We have added an explanation of this notation near the beginning of Section 5.1 on p. 24:

> Here, $\boldsymbol{X}_j \in \mathbb{R}$ denotes the jth element of the vector $\boldsymbol{X} \in \mathbb{R}^m$ and $\boldsymbol{X}_{-j} \in \mathbb{R}^{m-1}$ denotes all elements except the jth. Also, $X_{i,\bullet} \in \mathbb{R}^n$, $X_{\bullet,j} \in \mathbb{R}^m$, and $X_{\bullet,-j} \in \mathbb{R}^{n \times (m-1)}$ denote the ith row, jth column, and all columns but the jth of the matrix $X \in \mathbb{R}^{n \times m}$.