Submitted to the Annals of Statistics

RECONCILING MODEL-X AND DOUBLY ROBUST APPROACHES TO CONDITIONAL INDEPENDENCE TESTING

BY ZIANG NIU*[1,a], ABHINAV CHAKRABORTY*[1,b], OLIVER DUKES[2,d] AND EUGENE KATSEVICH[1,c]

[1]*Department of Statistics and Data Science, University of Pennsylvania ,* [a]*ziangniu@wharton.upenn.edu;* [b]*abch@wharton.upenn.edu;* [c]*ekatsevi@wharton.upenn.edu*

[2]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University,* [d]*oliver.dukes@ugent.be*

Model-X approaches to testing conditional independence between a predictor and an outcome variable given a vector of covariates usually assume exact knowledge of the conditional distribution of the predictor given the covariates. Nevertheless, model-X methodologies are often deployed with this conditional distribution learned in sample. We investigate the consequences of this choice through the lens of the distilled conditional randomization test (dCRT). We find that Type-I error control is still possible, but only if the mean of the outcome variable given the covariates is estimated well enough. This demonstrates that the dCRT is doubly robust, and motivates a comparison to the generalized covariance measure (GCM) test, another doubly robust conditional independence test. We prove that these two tests are asymptotically equivalent, and show that the GCM test is optimal against (generalized) partially linear alternatives by leveraging semiparametric efficiency theory. In an extensive simulation study, we compare the dCRT to the GCM test. These two tests have broadly similar Type-I error and power, though dCRT can have somewhat better Type-I error control but somewhat worse power in small samples or when the response is discrete. We also find that post-lasso based test statistics (as compared to lasso based statistics) can dramatically improve Type-I error control for both methods.

1. Introduction.

1.1. *Conditional independence testing and the model-X assumption.* Given a predictor $\boldsymbol{X} \in \mathbb{R}$, response $\boldsymbol{Y} \in \mathbb{R}$, and high-dimensional covariate vector $\boldsymbol{Z} \in \mathbb{R}^p$ drawn from a joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}_n$ (potentially varying with n to accommodate growing p), consider testing the hypothesis of conditional independence (CI)

$$H_{0n} : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z} \tag{1}$$

at level $\alpha \in (0, 1)$ using n data points

$$(X, Y, Z) \equiv \{(X_i, Y_i, Z_i)\}_{i=1,\ldots,n} \overset{\text{i.i.d.}}{\sim} \mathcal{L}_n. \tag{2}$$

Throughout the paper, boldface (respectively, regular) font indicates population (respectively, sample) quantities. In a high-dimensional regression setting, H_{0n} is a model-agnostic way of formulating the null hypothesis that predictor $\boldsymbol{X}$ is unimportant in the regression of $\boldsymbol{Y}$ on $(\boldsymbol{X}, \boldsymbol{Z})$ (Candès et al., 2018). In a causal inference setting with treatment $\boldsymbol{X}$, outcome $\boldsymbol{Y}$,

MSC2020 subject classifications: Primary 62J07, 62G10, 62G09.

Keywords and phrases: Model-X, Conditional randomization test, Conditional independence testing, Double robustness.

*These authors contributed equally to this work.

observed confounders $\boldsymbol{Z}$, and no unobserved confounders, H_{0n} is the null hypothesis of no causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (Pearl, 2009).

As Shah and Peters (2020) showed, the CI null hypothesis is too large in the sense that any test controlling Type-I error on H_{0n} must be powerless against all alternatives (if we assume, for example, that $\boldsymbol{Z}$ is continuously distributed). Therefore, additional assumptions must be placed on $\mathcal{L}_n$ to make progress. One such assumption is the *model-X (MX) assumption* (Candès et al., 2018), which states that $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is known exactly. Under the MX assumption, Candès et al. (2018) propose the MX knockoffs and conditional randomization test (CRT) methodologies, which have elegant finite-sample Type-I error control guarantees. These MX methodologies have since exploded in popularity, undergoing active methodological development and deployment in a range of applications.

One of the primary challenges in the practical application of MX methods is to obtain the required conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. Outside the context of randomized controlled experiments (Aufiero and Janson, 2022; Ham, Imai and Janson, 2022), the MX assumption is an approximation (Barber, Candès and Samworth, 2020; Huang and Janson, 2020; Li and Liu, 2022). In genome-wide association studies, a realistic parametric distribution can be postulated for this conditional law (Sesia, Sabatti and Candès, 2019), but the parameters of this distribution must still be learned from data. In practice, the conditional law is usually fit in sample on the same data that is used for testing, and then treated as if it were known (Candès et al., 2018; Sesia, Sabatti and Candès, 2019; Sesia et al., 2020; Bates et al., 2020; Liu et al., 2022; Li et al., 2021; Sesia et al., 2021; Barry et al., 2021). Such adaptations of MX methodologies are widely deployed, but their robustness and power properties have not been thoroughly investigated.

1.2. *Our contributions.* In this paper, we address this gap by investigating the properties of MX methods with $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ learned in sample. This investigation leads us to establish close connections between these methods and double regression approaches to CI testing, and to explore the optimality of CI tests against semiparametric alternatives. We focus our analyses on the *distilled conditional randomization test* (*dCRT*), a fast and powerful instance of the CRT (Liu et al., 2022), and the *generalized covariance measure (GCM) test*, a prototypical double regression approach to CI testing (Shah and Peters, 2020). Both tests involve learning $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ in sample. Our main contributions are outlined next:

1. **The dCRT with $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ learned in sample can have poor Type-I error control if $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ is learned poorly.** If $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z})$ is known exactly, then the dCRT has finite-sample Type-I error control regardless of $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{Z})$ or the quality of its estimate. This is no longer the case once $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z})$ is fit in sample, as we demonstrate in a numerical simulation and a theoretical counterexample (Section 3).
2. **The dCRT is doubly robust, in the sense that errors in $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ can be compensated for by better approximations of $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$.** The MX assumption shifts the modeling burden entirely from $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ to $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$. When the latter is fit in sample, shifting the modeling burden partially back towards $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ helps recover asymptotic Type-I error control, as we demonstrate theoretically (Section 4.2).
3. **The dCRT resampling distribution approaches normality, making this test asymptotically equivalent to the GCM test.** The dCRT is a resampling-based test, whereas the GCM test is asymptotic. In large samples, however, the resampling-based null distribution of the former converges to the $N(0,1)$ null distribution of the latter (Section 2). We show that these two tests are asymptotically equivalent against local alternatives (Section 4.1).
4. **The GCM test is asymptotically uniformly most powerful against local non-interacting alternatives.** Optimality results are widely prevalent in the semiparametric

literature, but not in the CI testing literature. We leverage semiparametric optimality theory to prove that the GCM is the optimal CI test against local (generalized) partially linear alternatives (Section 5), a broad class of alternatives in which $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not interact.

5. **In finite samples, the dCRT and GCM test have broadly similar Type-I error and power, with some exceptions.** The asymptotic equivalence between the dCRT and GCM test largely carries over to finite samples, as we demonstrate in numerical simulations (Section 6). The two tests have broadly similar Type-I error and power, although there is some divergence in small samples or when $\boldsymbol{Y}$ is discrete: in these cases dCRT can have somewhat better Type-I error control but somewhat worse power.
6. **In finite samples, replacing the lasso with the post-lasso markedly improves Type-I error control for both dCRT and GCM test.** In MX applications, the lasso is perhaps the most common approach for learning both $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$. However, we demonstrate in numerical simulations (Section 6) that the bias reduction offered by the post-lasso greatly improves Type-I error control in the context of both GCM test and dCRT, though at some cost in power.

On the way to making the aforementioned primary contributions, we make a few secondary contributions of independent interest:

7. **We reexamine numerical simulation setups from prior MX papers, finding that many have only low levels of marginal dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$.** Prior works have used numerical simulations to establish that MX methods are fairly robust when fitting $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ in sample. However, we note that the conditional independence testing problem (1) is difficult to the extent that $\boldsymbol{Z}$ induces spurious marginal dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (a "confounding" effect). We find simulation setups in prior works have low levels of this marginal dependence (Section 6.1), potentially leading to optimistic conclusions.
8. **We collate a number of conditional analogs of classical convergence theorems (some but not all novel).** The dCRT involves resampling conditionally on the observed data, so its asymptotic analysis requires reasoning about convergence after conditioning on a σ-algebra that changes with n. We state and prove conditional analogs of Slutsky's theorem, the law of large numbers, the central limit theorem, and other classical convergence theorems (Appendix B of the Supplementary Material (Niu et al., 2023)). These results are not surprising, but at least some appear novel.
9. **We prove a sharpened theorem on optimality in semiparametric testing.** In the literature on semiparametric *estimation*, an estimator need only be regular *in the vicinity of a point* for efficiency bounds to hold, whereas popular textbooks (Van Der Vaart, 1998; Kosorok, 2008) state semiparametric *testing* optimality results *globally*: a test must control Type-I error on the entire semiparametric null, rather than just in the vicinity of a point, for efficiency bounds to hold. We address this gap by proving a stronger local optimality result for semiparametric testing (Appendix E.1 in Niu et al. (2023)).

1.3. *Related work.* We split related works into three categories: those investigating the robustness of the original MX methods (knockoffs and CRT) to misspecification of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$, those proposing new variants of MX methods designed for robustness, and those investigating the power of MX methods.

Robustness of original MX methods. One line of work investigates the Type-I error of knockoffs (Barber, Candès and Samworth, 2020) and the CRT (Berrett et al., 2020) when $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is misspecified, in the worst case over all possible test statistics and all possible distributions $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$. In the context of the CRT, Berrett et al. (2020) proved that the excess Type-I error is upper-bounded by the total variation error in approximating $\prod_{i=1}^n \mathcal{L}_n(X_i \mid Z_i)$, alongside a matching lower bound. A similar style of result holds for knockoffs (Barber,

Candès and Samworth, 2020). These works do not allow for $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ to be fit in sample, however. Even if they applied in this case, one could at most hope for the aforementioned TV distance to be $O(1)$. For these worst-case bounds to guarantee asymptotic Type-I error control, one would need to learn the conditional distribution $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ on an additional unlabeled sample of size $N \gg n$. For specific test statistics, however, MX methods may be more robust. For example, Katsevich and Ramdas (2022) proved that the distilled CRT (an instance of the CRT with a product-of-residuals test statistic) has asymptotic Type-I error control when only the first two moments of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ are correct but higher-order moments may be misspecified. Even this weaker assumption cannot be expected to hold when $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is fit in sample, however. Another line of work (Fan et al., 2020a,b; Fan, Gao and Lv, 2023) probes the robustness of MX knockoffs with imperfect covariate distribution for a variety of specific test statistics and covariate distributions, with Fan et al. (2020a); Fan, Gao and Lv (2023) allowing for the covariate distribution to be learned in sample while guaranteeing asymptotic FDR control. The robustness aspects of the present work can be viewed as complementing the latter two existing works; we focus on the CRT rather than on knockoffs.

New variants of MX methods designed for robustness. Modifications of the originally proposed CRT and knockoffs have been designed specifically to have improved robustness to misspecifications of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$. For example, if this law is known to belong to a parametric family with a low-dimensional sufficient statistic, a variant of MX knockoffs can be carried out conditionally on this sufficient statistic without needing to accurately estimate the parameters themselves (Huang and Janson, 2020). The former methodology enjoys a double robustness property, related to but different from the one we state for the dCRT (see contribution 2). Even in the absence of a low-dimensional sufficient statistic, Barber and Janson (2022) proposed a variant of the CRT based on conditioning on an approximate sufficient statistic. Another method, the *conditional permutation test* (Berrett et al., 2020), is a variant of the CRT based on conditioning on the order statistics of $\{X_i\}$ rather than on a sufficient statistic for $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$. This test was shown to be more robust than the CRT to misspecification of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$. Finally, the *Maxway CRT* (Li and Liu, 2022) has recently been proposed as a doubly robust variant of the dCRT. In this manuscript, we argue that, in fact, the dCRT itself is doubly robust. Overall, our goal is not to introduce new methodology but to study the robustness of (a special case of) the originally proposed CRT. Despite the emergence of several new variants of MX methods like those described above, the originally proposed CRT and knockoffs remain the most widely deployed MX methods in practice.

Power of MX methods. A number of works have investigated the power of the CRT and knockoffs (Weinstein, Barber and Candes, 2017; Liu and Rigollet, 2019; Weinstein et al., 2020; Fan et al., 2020a,b; Katsevich and Ramdas, 2022; Wang and Janson, 2022; Spector and Fithian, 2022), although only Fan et al. (2020a,b); Katsevich and Ramdas (2022) do not assume that $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is known exactly (the MX assumption). Beyond calculating power against certain alternatives, Katsevich and Ramdas (2022) and Spector and Fithian (2022) discuss test statistic choices for the CRT and MX knockoffs, respectively, that yield optimal power under the MX assumption. In the current work, we investigate not just optimal *statistics* for certain methods but optimal CI *methods* against certain classes of alternatives, and without assuming that $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is known (see contribution 4). We defer the discussion of further optimality-related work to Section 5.3.

1.4. *Preliminaries: The dCRT and GCM tests.* Here we formally define two of the primary CI tests under investigation, the dCRT and the GCM test. For both of these, it will be useful to define

$$\mu_{n,x}(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \quad \text{and} \quad \mu_{n,y}(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]. \tag{3}$$

1.4.1. *The dCRT and* $\widehat{\text{dCRT}}$. A simple approach to CI testing under the MX assumption is the *conditional randomization test* (CRT, Candès et al. (2018)), which controls Type-I error not just asymptotically (28) but in finite samples as well. The CRT is based on constructing a null distribution for any test statistic $T_n(X, Y, Z)$ by resampling X conditionally on Z using the known conditional law $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ (Algorithm 1).

Algorithm 1: The conditional randomization test (CRT).

Input: Data (X, Y, Z), number of randomizations M, conditional law $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$.

Compute $T_n(X, Y, Z)$;

for $m = 1, 2, \ldots, M$ **do**

  Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^n \mathcal{L}_n(X_i|Z_i)$ and compute $T_n(\widetilde{X}^{(m)}, Y, Z)$;

end

Output: CRT p-value $\frac{1}{M+1}(1 + \sum_{m=1}^M \mathbb{1}\{T_n(\widetilde{X}^{(m)}, Y, Z) \geq T_n(X, Y, Z)\})$.

The test statistic T_n is usually a measure of variable importance for the predictor X based on a predictive model of Y on (X, Z) trained on the given data. In general, the CRT requires retraining this predictive model for each resampled dataset $(\widetilde{X}^{(m)}, Y, Z)$, and can therefore be computationally costly.

Motivated by the high computational cost of the CRT, a faster but similarly powerful *distilled CRT* (dCRT, Liu et al. (2022)) was proposed as a special case based on a test statistic of the form

$$T_n^{\text{dCRT}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)).$$

Here, $\mu_{n,x}$ is known under the MX assumption and $\widehat{\mu}_{n,y}$ is trained in sample. The dCRT is fast because it does not require retraining the predictive model $\widehat{\mu}_{n,y}$ for each resampled dataset, as it depends on (Y, Z) only. Variants of the dCRT have now been deployed in genetics (Bates et al., 2020) and genomics (Barry et al., 2021) applications. As discussed in Section 1.1, MX methodologies (including the dCRT) are usually deployed by learning $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ in sample. For clarity, we give the dCRT with $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ fit in sample a new name: $\widehat{\text{dCRT}}$. This procedure is based on the test statistic

$$T_n^{\widehat{\text{dCRT}}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)), \tag{4}$$

where $\widehat{\mu}_{n,x}(Z_i) \equiv \mathbb{E}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i]$. The $\widehat{\text{dCRT}}$ procedure is outlined in Algorithm 2; one of the primary goals of this paper is to study this procedure.

Algorithm 2: The $\widehat{\text{dCRT}}$.

Input: Data (X, Y, Z), number of randomizations M.

Learn $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ based on (X, Z) and $\widehat{\mu}_{n,y}(\boldsymbol{Z})$ based on (Y, Z);

Compute $T_n^{\widehat{\text{dCRT}}}(X, Y, Z)$;

for $m = 1, 2, \ldots, M$ **do**

  Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^n \widehat{\mathcal{L}}_n(X_i|Z_i)$ and compute

$$T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (\widetilde{X}_i^{(m)} - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)); \tag{5}$$

end

Output: $\widehat{\text{dCRT}}$ p-value $\frac{1}{M+1}(1 + \sum_{m=1}^M \mathbb{1}\{T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z) \geq T_n^{\widehat{\text{dCRT}}}(X, Y, Z)\})$.

The resampled test statistics $T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z)$ (5) have four arguments instead of three in order to emphasize that the conditional mean $\widehat{\mu}_{n,x}(\cdot)$ is not refit upon resampling.

1.4.2. *The GCM test and double robustness.* Another CI test is the GCM test (Shah and Peters, 2020), defined as

$$\phi_n^{\text{GCM}}(X, Y, Z) \equiv \mathbb{1}(T_n^{\text{GCM}}(X, Y, Z) > z_{1-\alpha}), \tag{6}$$

where

(7)

$$T_n^{\text{GCM}}(X, Y, Z) \equiv \frac{1}{\widehat{S}_n^{\text{GCM}}} \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)) \equiv \frac{1}{\widehat{S}_n^{\text{GCM}}} T_n^{\widehat{\text{dCRT}}}(X, Y, Z)$$

and $(\widehat{S}_n^{\text{GCM}})^2$ is the empirical variance of the product-of-residual summands:

$$(\widehat{S}_n^{\text{GCM}})^2 \equiv \widehat{\text{Var}}\{(X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))\}. \tag{8}$$

It controls Type-I error if the following in-sample mean-squared error quantities are small (Shah and Peters, 2020):

$$E_{n,x} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2\right)^{1/2}; \; E'_{n,x} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \text{Var}_{\mathcal{L}_n}[Y_i|Z_i]\right)^{1/2};$$

$$E_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2\right)^{1/2}; \; E'_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2 \text{Var}_{\mathcal{L}_n}[X_i|Z_i]\right)^{1/2}.$$

In particular, Shah and Peters (2020) require that

$$E_{n,x}E_{n,y} = o_{\mathcal{L}_n}(n^{-1/2}), \; E'_{n,x} = o_{\mathcal{L}_n}(1), \; E'_{n,y} = o_{\mathcal{L}_n}(1), \tag{SP1}$$

and, for some constants $c_1, c_2, \delta > 0$,

$$\begin{aligned} &\inf_n \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))^2(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))^2] > c_1 \\ &\sup_n \mathbb{E}_{\mathcal{L}_n}[|(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))|^{2+\delta}] < c_2. \end{aligned} \tag{SP2}$$

The GCM test is therefore *doubly robust* in the sense that it controls Type-I error if the product of the estimation errors for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ ($E_{n,x}E_{n,y}$) converges to zero at

the $o_{\mathcal{L}_n}(n^{-1/2})$ rate. Note that this is a *rate double robustness* property rather than a *model double robustness* property; see Smucler, Rotnitzky and Robins (2019) for a discussion of this distinction. Unless otherwise specified, we use the term "doubly robust" to refer to rate double robustness of Type-I error control.

2. $\widehat{\text{dCRT}}$ resampling distribution converges to normal. To make it easier to analyze the asymptotic properties of the $\widehat{\text{dCRT}}$, in this section we prove that it is asymptotically equivalent to the resampling-free $\widehat{\text{MX}(2)}$ F-test, a variant of the MX(2) F-test (Katsevich and Ramdas, 2022) where the first two moments of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ are estimated in sample. This equivalence was already shown by these authors in the case when $\mu_{n,x}$ is known and $\widehat{\mu}_{n,y}$ is fit out of sample (see their Theorem 2). They conjectured that the equivalence continues to hold when $\widehat{\mu}_{n,y}$ is fit in sample. Here, we prove this conjecture, not just when $\widehat{\mu}_{n,y}$ is fit in sample, but also when the first two moments of $\mu_{n,x}$ are unknown and also fit in sample.

Note that the variance of the resampling distribution of $T_n^{\widehat{\text{dCRT}}}$ is

$$(\widehat{S}_n^{\widehat{\text{dCRT}}})^2 \equiv \text{Var}_{\widehat{\mathcal{L}}_n}[T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z] = \frac{1}{n}\sum_{i=1}^{n} \text{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2. \tag{9}$$

It will be convenient to reformulate $\widehat{\text{dCRT}}$ as

$$\begin{aligned}
\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) &\equiv \mathbb{1}(T_n^{\widehat{\text{dCRT}}}(X,Y,Z) > \mathbb{Q}_{1-\alpha}[T_n^{\widehat{\text{dCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z]) \\
&= \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(X,Y,Z) > \mathbb{Q}_{1-\alpha}\left[\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z\right]\right) \\
&\equiv \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(X,Y,Z) > C_n^{\widehat{\text{dCRT}}}(X,Y,Z)\right).
\end{aligned}$$

The conditional $1-\alpha$ quantile $C_n^{\widehat{\text{dCRT}}}(X,Y,Z)$ is defined in the last line above. Note that this test is obtained from that in Algorithm 2 by sending $M \to \infty$; we focus our theoretical analysis here and throughout on this infinite-resamples limit of the $\widehat{\text{dCRT}}$. Here, the α conditional quantile $\mathbb{Q}_\alpha[W \mid \mathcal{F}]$ of a random variable W given a σ-algebra $\mathcal{F}$ is defined via

$$\mathbb{Q}_\alpha[W \mid \mathcal{F}] \equiv \inf\{t : \mathbb{P}[W \le t \mid \mathcal{F}] \ge \alpha\}. \tag{10}$$

One would expect, based on the central limit theorem, that the conditional distribution of the ratio $T_n^{\widehat{\text{dCRT}}}(\widetilde{X},X,Y,Z)/\widehat{S}_n^{\widehat{\text{dCRT}}}$ tends to $N(0,1)$. This statement is complicated by the conditioning event, which requires us to be careful to define conditional convergence in distribution:

DEFINITION 1. For each n, let W_n be a random variable and let $\mathcal{F}_n$ be a σ-algebra. Then, we say W_n converges in distribution to a random variable W conditionally on $\mathcal{F}_n$ if

$$\mathbb{P}[W_n \le t \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{P}[W \le t] \text{ for each } t \in \mathbb{R} \text{ at which } t \mapsto \mathbb{P}[W \le t] \text{ is continuous.} \tag{11}$$

We denote this relation via $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$.

Based on an extension of the Lyapunov central limit theorem to conditional convergence in distribution (Theorem 5 in Niu et al. (2023)), we get the following result:

THEOREM 1. *Suppose the sequences of true and learned laws $\mathcal{L}_n$ and $\widehat{\mathcal{L}}_n$ satisfy the following two nondegeneracy properties:*

(NDG1) $$\mathbb{P}_{\mathcal{L}_n}[(\widehat{S_n^{\text{dCRT}}})^2 \geq \epsilon] \to 1 \text{ for some } \epsilon > 0;$$

(NDG2) $$0 < \text{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i], (Y_i - \widehat{\mu}_{n,y}(Z_i))^2, (Y_i - \mu_{n,y}(Z_i))^2 < \infty \text{ almost surely.}$$

If the conditional Lyapunov condition

(Lyap-1) $$\frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}\left[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z\right] \xrightarrow{p} 0$$

is satisfied for some $\delta > 0$, then

(12) $$\frac{1}{\widehat{S_n^{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z \xrightarrow{d,p} N(0,1)$$

and therefore

(13) $$C_n^{\widehat{\text{dCRT}}}(X, Y, Z) \equiv \mathbb{Q}_{1-\alpha}\left[\frac{1}{\widehat{S_n^{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z\right] \xrightarrow{p} z_{1-\alpha}.$$

This suggests that the $\widehat{\text{dCRT}}$ is asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test, defined

(14) $$\phi_n^{\widehat{\text{MX(2)}}}(X, Y, Z) \equiv \mathbb{1}\left(\frac{1}{\widehat{S_n^{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > z_{1-\alpha}\right).$$

Indeed, we have the following corollary.

COROLLARY 1. *Consider a sequence of laws $\mathcal{L}_n$ satisfying the assumptions* (NDG1)*,* (NDG2)*, and* (Lyap-1) *of Theorem 1, and assume that the test statistic does not accumulate near $z_{1-\alpha}$, i.e.*

(15) $$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n^{\widehat{\text{dCRT}}}(X, Y, Z) - z_{1-\alpha}| \leq \delta] = 0.$$

Then, the $\widehat{\text{dCRT}}$ is asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test:

(16) $$\lim_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z) = \phi_n^{\widehat{\text{MX(2)}}}(X, Y, Z)] = 1.$$

This result extends Katsevich and Ramdas (2022, Theorem 2) by allowing $\widehat{\mu}_{n,x}$ and $\widehat{\mu}_{n,y}$ to be fit in sample, rather than assuming $\mu_{n,x}$ is known and $\widehat{\mu}_{n,y}$ is fit out of sample. It is a first indication that the $\widehat{\text{dCRT}}$ approximates a test based on asymptotic normality.

3. $\widehat{\text{dCRT}}$ is not robust for general $\widehat{\mu}_{n,y}$. One of the hallmarks of MX inference is that it requires "no restriction on the dimensionality of the data or the conditional distribution of $[\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})]$" (Candès et al., 2018). For the CRT, this means that Type-I error is controlled in finite samples, regardless of the test statistic used or the distribution of the response variable. If $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is described by a parametric model with k unknown parameters and we have $N \gg n \cdot k$ unlabeled samples to learn this model, then at least asymptotic Type-I error control is still possible without assumptions on $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$ (Berrett et al., 2020). By contrast, in this section we show that when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is approximated in sample, we cannot expect Type-I error control without assumptions on the response variable.

Let us consider a simple null model $\mathcal{L}_n$ with

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, I_p), \quad \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1), \quad \text{and} \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1). \tag{17}$$

Suppose we fit $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ via a ridge regression while using the trivial estimate $\widehat{\mu}_{n,y}(\boldsymbol{Z}) \equiv 0$ for $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$. To build intuition while avoiding technical difficulties, we loosely approximate the ridge regression estimator as $\widehat{\beta}_n \equiv (1 - \frac{c}{\sqrt{n}})\beta$, where the $1/\sqrt{n}$ error term reflects that we are fitting $\widehat{\beta}_n$ in sample (and is optimistic in the sense that it ignores possible growth in p). Then, consider the $\widehat{\text{dCRT}}$ based on $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\widehat{\beta}_n, 1)$ and $\widehat{\mu}_{n,y}(\boldsymbol{Z}) \equiv 0$. In this case, the normality of $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ leads to normality of the resampling distribution holding not just asymptotically (12) but in finite samples as well. Therefore, the $\widehat{\text{dCRT}}$ is equal to the $\widehat{\text{MX}(2)}$ F-test:

$$\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z) = \mathbb{1}\left(\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}} \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - Z_i^T\widehat{\beta}_n)Y_i > z_{1-\alpha}\right). \tag{18}$$

On the other hand, it is easy to derive that

$$\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}} \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - Z_i^T\widehat{\beta}_n)Y_i \xrightarrow{d} N\left(\frac{c\|\beta\|^2}{\sqrt{\|\beta\|^2 + 1}}, 1\right). \tag{19}$$

Therefore, the limiting Type-I error of the $\widehat{\text{dCRT}}$ in this case is

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z)] = 1 - \Phi\left(z_{1-\alpha} - \frac{c\|\beta\|^2}{\sqrt{\|\beta\|^2 + 1}}\right), \tag{20}$$

which can be made arbitrarily close to one as $c \to \infty$. This issue is caused by a combination of the $O(1/\sqrt{n})$ shrinkage bias in the estimator for $\mu_{n,x}$ and the failure to estimate $\mu_{n,y}$. This leaves an $O(1/\sqrt{n})$ correlation between $\boldsymbol{X} - \widehat{\mu}_{n,x}(\boldsymbol{Z})$ and $\boldsymbol{Y}$ induced by $\boldsymbol{Z}$, which shifts the mean of the null distribution of the $\widehat{\text{dCRT}}$ test statistic away from zero by a nontrivial amount.

Numerical simulations (although with lasso instead of ridge regression) confirm this phenomenon. We constructed a numerical simulation based on the null model (17) with $n = 1600$, $p = 400$, and β having only $s = 5$ nonzero entries (see Section 6.2 below for more on our data-generating model). In this setting, we applied the $\widehat{\text{dCRT}}$ using the cross-validated lasso and intercept-only models to estimate $\mu_{n,x}$ and $\mu_{n,y}$, respectively. As we increased the magnitude of the coefficient vector β, this test exhibited significant loss of Type-I error control (Figure 1). By contrast, using the lasso instead of the intercept-only model to estimate $\mu_{n,y}$ reduced the Type-I error to nearly the nominal level.

So even when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is estimated at a parametric rate (albeit with regularization), the $\widehat{\text{dCRT}}$ can have inflated Type-I error rate for certain test statistics. A similar observation was made by Li and Liu (2022) (see the discussion after Theorem 3). Similar phenomena have been noted in the contexts of causal inference (Dukes and Vansteelandt, 2020) and doubly robust estimation (Chernozhukov et al., 2018, 2022); in the latter literature this issue is called "regularization bias." We note that poor estimation of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$, in conjunction with the plug-in resampling scheme of the $\widehat{\text{dCRT}}$ can also lead to conservative inference rather than liberal inference. This happens in cases when $\widehat{\beta}_n$ is an efficient estimator of β, e.g. that derived from ordinary least squares. In the causal inference context, this conservatism is a consequence of the fact that using estimated propensity scores can lead to more efficient estimates than using known propensity scores (Robins, Mark and Newey, 1992; Henmi and Eguchi, 2004). If the

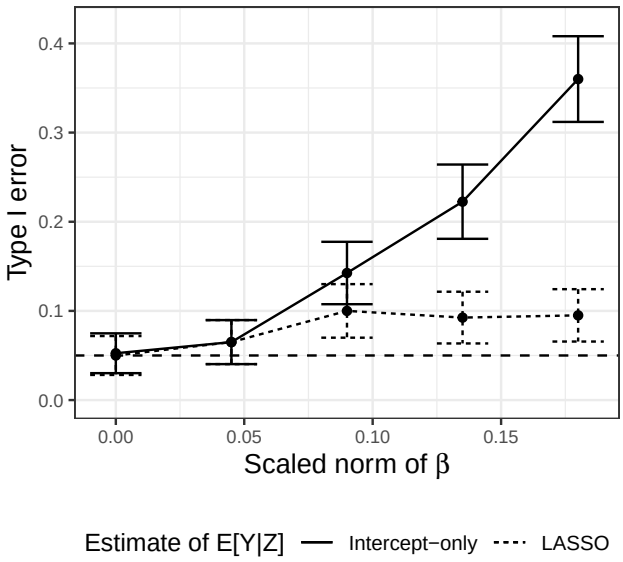

Fig 1: The Type-I error of two instances of the $\widehat{\text{dCRT}}$ under the data-generating model (17), depending on which method is used to estimate $\mu_{n,y}$, when the lasso is used to estimate $\mu_{n,x}$. Improved estimation of $\mu_{n,y}$ leads to markedly reduced Type-I error.

propensity score is estimated but the standard error is constructed as though it were known, then conservative inference would result.

As already alluded to, the Type-I error inflation in the above example stems from the fact that

$$\mathbb{E}_{\mathcal{L}_n}[(\widehat{\mu}_{n,x}(\boldsymbol{Z}) - \mu_{n,x}(\boldsymbol{Z}))(\widehat{\mu}_{n,y}(\boldsymbol{Z}) - \mu_{n,y}(\boldsymbol{Z}))] = O(1/\sqrt{n}),$$

a rate insufficient for Type-I error control. If we had at least consistency of $\widehat{\mu}_{n,y}(\boldsymbol{Z})$, then this rate would improve to $o(1/\sqrt{n})$ and Type-I error control would be restored. This intuition is supported by the simulation results in Figure 1, where estimating $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ via lasso brought the Type-I error down to nearly the nominal level. This discussion suggests that, if $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is learned in sample (or on an external sample of similar size), then assumptions must be placed not only on $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ but also on $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$ for Type-I error control. This motivates us to investigate the double robustness of the $\widehat{\text{dCRT}}$ and compare it to the GCM test.

4. $\widehat{\text{dCRT}}$ is doubly robust and equivalent to GCM test. Of course, in practice $\widehat{\mu}_{n,y}$ is not fit as naively as in the counterexample from Section 3. The conditional mean $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ is usually approximated via a machine learning algorithm, as improved approximation of this quantity improves the power of the dCRT (Katsevich and Ramdas, 2022). In the context where $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ must be approximated, we claim that more accurate estimation of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ can improve not just the power but also the Type-I error control of the $\widehat{\text{dCRT}}$. We formalize this by showing that the $\widehat{\text{dCRT}}$ is *doubly robust* (recall Section 1.4). This property is a consequence of the fact that, under the null, the $\widehat{\text{dCRT}}$ is asymptotically equivalent to the GCM test, which itself is doubly robust. This equivalence also implies that the $\widehat{\text{dCRT}}$ and GCM test have the same asymptotic power against contiguous alternatives.

4.1. *Equivalence between GCM test and* $\widehat{\text{dCRT}}$. When comparing the GCM test (6) to the $\widehat{\text{MX}(2)}$ F-test (14), which is asymptotically equivalent to the $\widehat{\text{dCRT}}$ (Corollary 1), the

only difference is the normalization term. Under the null hypothesis, this difference vanishes asymptotically as long as the estimated variance $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ is consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i] - \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i])\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0. \tag{21}$$

In preparation to state our equivalence result, we augment the assumption (SP1) as follows:

$$E_{n,x}E_{n,y} = o_{\mathcal{L}_n}(n^{-1/2}),\ E'_{n,x} = o_{\mathcal{L}_n}(1),\ E'_{n,y} = o_{\mathcal{L}_n}(1),\ \widehat{E}'_{n,y} = o_{\mathcal{L}_n}(1), \tag{SP1'}$$

where

$$\widehat{E}'_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i]\right)^{1/2}. \tag{22}$$

Furthermore, we denote by

$$\mathscr{L}_n^0 \equiv \{\mathcal{L}_n : \mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = \mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z}) \times \mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})\} \tag{23}$$

the set of laws satisfying conditional independence.

THEOREM 2. *Suppose $\mathcal{L}_n \in \mathscr{L}_n^0$ is a sequence of laws satisfying the assumptions* (SP1') *and* (SP2)*, the nondegeneracy condition* (NDG2)*, the variance consistency property* (21) *and the Lyapunov condition*

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}\left[|Y_i - \mu_{n,y}(Z_i)|^{2+\delta} \mid Z_i\right]\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z] \xrightarrow{p} 0. \tag{Lyap-2}$$

Then, the $\widehat{\text{dCRT}}$ *and GCM variance estimates are asymptotically equivalent:*

$$\frac{(\widehat{S}_n^{\widehat{\text{dCRT}}})^2}{(\widehat{S}_n^{\text{GCM}})^2} \xrightarrow{p} 1, \tag{24}$$

as are the $\widehat{\text{dCRT}}$ *and GCM tests themselves:*

$$\lim_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z) = \phi_n^{\text{GCM}}(X, Y, Z)] = 1. \tag{25}$$

The variance consistency property (21) is relatively easy to achieve, given the other assumptions of Theorem 2. The following proposition states two sufficient conditions for this property.

PROPOSITION 1. *If the assumptions of Theorem 2 other than variance consistency* (21) *hold, then the latter property holds in the following two cases:*

1. $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i] \equiv (X_i - \widehat{\mu}_{n,x}(Z_i))^2$*;*
2. $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X}|\boldsymbol{Z}] \equiv f(\widehat{\mu}_{n,x}(\boldsymbol{Z}))$*, if*
   - $\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = f(\mu_{n,x}(\boldsymbol{Z}))$ *for f Lipschitz on domain* $\cup_{n=1}^{\infty}\mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$ *and* $\mathrm{supp}(\widehat{\mu}_{n,x}(\boldsymbol{Z})) \subseteq \mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$ *almost surely for every n;*
   - $\sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty$ *for some $\delta > 0$.*

The first variance estimate given in the proposition can always be applied; the second applies to cases when the mean-variance relationship for $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is known and Lipschitz on the convex hull of the support of $\boldsymbol{X}$, denoted $\mathrm{Conv}(\mathcal{L}_n(\boldsymbol{X}))$. This is the case, for example, if $\boldsymbol{X}$ is binary and we define $f(t) \equiv t(1-t)$.

One consequence of Theorem 2 is that the $\widehat{\text{dCRT}}$ and GCM test are also asymptotically equivalent against local alternatives, so in particular have the same power.

COROLLARY 2. *If $\mathcal{L}'_n$ is a sequence of alternative distributions that is contiguous to a sequence $\mathcal{L}_n \in \mathscr{L}_n^0$ satisfying the assumptions of Theorem 2, then the* $\widehat{\text{dCRT}}$ *and GCM tests are asymptotically equivalent against $\mathcal{L}'_n$:*

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}'_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) = \phi_n^{\text{GCM}}(X,Y,Z)] = 1 \tag{26}$$

and therefore have the same asymptotic power:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}'_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}'_n}[\phi_n^{\text{GCM}}(X,Y,Z)] = 0. \tag{27}$$

By constructing a null distribution via resampling, the CRT allows for arbitrarily complicated test statistics whose asymptotic distributions are not known. For the $\widehat{\text{dCRT}}$, however, the resampling-based null distribution simply recapitulates the asymptotic normal distribution used by the GCM test (Theorems 1 and 2). Therefore, at least in large samples, the extra computational burden of resampling is unnecessary as the equivalent GCM can be applied instead.

4.2. *Double robustness of* $\widehat{\text{dCRT}}$. Another consequence of Theorem 2 is that the $\widehat{\text{dCRT}}$ is doubly robust under the variance consistency condition (21), since it is equivalent under the null hypothesis to the doubly robust GCM test. We will formulate this result in terms of a class of distributions $\mathscr{R}_n$ satisfying some regularity assumptions. For any regularity class $\mathscr{R}_n$, we consider testing the null hypothesis

$$H_{0n}(\mathscr{R}_n) : \mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n.$$

A sequence of tests $\phi_n : (X,Y,Z) \mapsto [0,1]$ of this null hypothesis has asymptotic Type-I error control if

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n(X,Y,Z)] \le \alpha. \tag{28}$$

COROLLARY 3. *Let $\mathscr{R}_n$ be a sequence of regularity conditions such that for any sequence $\mathcal{L}_n \in \mathscr{R}_n$, we have the nondegeneracy condition* (NDG2)*, the Lyapunov condition* (Lyap-2)*, the conditions* (SP1')*and* (SP2)*, and consistent variance estimates* (21)*. Then, the* $\widehat{\text{dCRT}}$ *has asymptotic Type-I error control over $\mathscr{L}_n^0 \cap \mathscr{R}_n$ in the sense of the definition* (28)*.*

Therefore, Type-I error control requires accuracy of only the first two moments of $\widehat{\mathcal{L}}_n$, in parallel to Theorem 2 of Katsevich and Ramdas (2022). The condition on the second moment of $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ is needed because the variance of the resampling distribution must not be smaller (asymptotically) than the true variance of the test statistic. This condition does not require much more than accurate estimation of the first moments (Proposition 1). It can be dropped altogether if we build normalization directly into the $\widehat{\text{dCRT}}$ test statistic. We explore this possibility in Appendix A in the supplementary material (Niu et al., 2023).

Our double robustness result for the dCRT evokes the double robustness result proved for a conditional variant of MX knockoffs by Huang and Janson (2020). We note that these two results refer to two different notions of double robustness. Corollary 3 states that the dCRT is *rate doubly robust*, while Huang and Janson (2020) finds that conditional knockoffs are *model doubly robust* (Smucler, Rotnitzky and Robins, 2019). Our result requires a condition on the product of the estimation rates for $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$, and accommodates high-dimensional settings. The double robustness of conditional knockoffs requires that one

of $\mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ belongs to a correctly specified, low-dimensional parametric family. We leave the investigation of the CRT's *model double robustness* to future work.

Our conclusion that $\widehat{\text{dCRT}}$ is doubly robust initially appears at odds with the statement that "the model-X CRT...does not pursue such double robustness through learning and adjusting for both $X|Z$ and $Y|Z$..." (Li and Liu, 2022). This statement is in reference to the worst-case performance of the CRT across all possible test statistics (Berrett et al., 2020). We agree that this worst-case performance can be poor when learning $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ in sample (Section 3). However, the test statistics applied in conjunction with the CRT (such as the dCRT statistic) do usually involve learning and adjusting for $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$. In this sense, practical applications of the (d)CRT do learn and adjust for both $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$; the former is learned when approximating the "model for X" and the latter when computing the test statistic. If the quality of these estimates is sufficiently good, then the $\widehat{\text{dCRT}}$ will control Type-I error (Corollary 3).

5. GCM test is optimal against certain alternatives. We have shown that, in large samples, the $\widehat{\text{dCRT}}$ has the same power against local alternatives as the resampling-free GCM test. Of course, other instances of the much more general CRT paradigm have better power than the GCM test against certain alternatives. We show in this section, however, that this is not the case for generalized partially linear models (GPLMs), a broad class of alternatives. In fact, the GCM test is asymptotically most powerful against GPLM alternatives. We leverage classical semiparametric efficiency theory (Choi, Hall and Schick, 1996; Van Der Vaart, 1998; Kosorok, 2008) to prove this result. We state our optimality result in Section 5.1, give an example of its application in Section 5.2, and then compare it to existing semiparametric optimality results in Section 5.3.

5.1. *Optimality result.* To facilitate the link with semiparametric theory, in this section of the paper we operate in a fixed-dimensional setting. Accordingly, we drop the subscript n from $\mathscr{L}_n^0$ and $\mathscr{R}_n$. For each value of n, we have $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{1+1+p}$ for fixed p. We will seek power against semiparametric GPLM alternatives of the form

$$\mathcal{L}_\theta(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \equiv \mathcal{L}_{\beta,\eta}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \equiv \mathcal{L}_{x,z}(\boldsymbol{X}, \boldsymbol{Z}) \times f_{\boldsymbol{\eta}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}), \quad \boldsymbol{\eta} = \boldsymbol{X}\beta + g(\boldsymbol{Z}). \tag{29}$$

Here, $\mathcal{L}_{x,z}$ is a fixed law, f_η is a one-parameter exponential family with natural parameter $\eta \in \mathbb{R}$ and log-partition function ψ, $\beta \in \mathbb{R}$ and

$$g \in \mathcal{H}_g \subseteq L^2(\mathcal{L}_{x,z}(\boldsymbol{Z})), \tag{30}$$

where $\mathcal{H}_g$ is a linear subspace of the L^2 space of functions on $\mathbb{R}^p$ with the measure $\mathcal{L}_{x,z}(\boldsymbol{Z})$. The alternatives (29) are those where $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ follows an exponential family distribution with natural parameter linear in $\boldsymbol{X}$ and potentially nonlinear in $\boldsymbol{Z}$. Note that GPLMs include linear and generalized linear models as special cases, and therefore cover a broad range of alternative distributions.

We focus on power against local alternatives $\mathcal{L}_{\theta_n(h)}$ near $\theta_0 \equiv (0, g_0)$, defined by

$$\theta_n(h) \equiv \theta_n(h_\beta, h_g) \equiv (h_\beta/\sqrt{n}, g_0 + h_g/\sqrt{n}), \quad \text{for} \quad h \equiv (h_\beta, h_g) \in (0, \infty) \times \mathcal{H}_g. \tag{31}$$

We leave the dependence of $\theta_n(h)$ on g_0 implicit. Next, we define asymptotic optimality against such local alternatives following Choi, Hall and Schick (1996):

DEFINITION 2. For $h \in (0, \infty) \times \mathcal{H}_g$, we say a test ϕ_n^* is the locally asymptotically most powerful level α test of

$$H_0 : \mathcal{L} \in \mathscr{R} \subseteq \mathscr{L}^0 \quad \text{versus} \quad H_{1n} : \mathcal{L} = \mathcal{L}_{\theta_n(h)} \tag{32}$$

if ϕ_n^* has asymptotic Type-I error control over $\mathscr{R}$ at level α and for any other test ϕ_n satisfying the same property we have

$$\limsup_{n\to\infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n(X,Y,Z)] \leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n^*(X,Y,Z)]. \tag{33}$$

If this is true for every $h \in (0,\infty) \times \mathcal{H}_g$, such a test is locally asymptotically uniformly most powerful at g_0, or LAUMP(g_0). A test is LAUMP($\mathcal{S}$) against $\mathcal{L}_{\theta_n(h)}$ for $h \in (0,\infty) \times \mathcal{H}_g$ if it is LAUMP(g_0) for each $g_0 \in \mathcal{S} \subseteq \mathcal{H}_g$.

Finally, define

$$s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]]. \tag{34}$$

We are now ready to state our main optimality result.

THEOREM 3. *Consider the conditional independence testing problem (32), with a collection of null distributions $\mathscr{R} \subseteq \mathscr{L}^0$ satisfying some regularity conditions, a linear subspace $\mathcal{H}_g \subseteq L^2(\mathcal{L}_{x,z}(\boldsymbol{Z}))$ specifying possible values for the nonparametric component g in the GPLM alternative model (29), and some subset $\mathcal{S} \subseteq \mathcal{H}_g$. If the following four assumptions hold:*

$$\textit{assumptions (SP1) and (SP2) hold for all } \mathcal{L} \in \mathscr{R}, \tag{35}$$

$$\ddot{\psi} = K > 0 \textit{ and } \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty \textit{ OR } \mathrm{supp}(\boldsymbol{X},\boldsymbol{Z}) \textit{ is compact and } \mathcal{H}_g \subseteq C(\mathbb{R}^p), \tag{36}$$

$$\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\,\cdot\,] \in \mathcal{H}_g, \tag{37}$$

$$\forall\ g_0 \in \mathcal{S}, h_g \in \mathcal{H}_g,\ \mathcal{L}_{\theta_n(0,h_g)} \in \mathscr{R} \textit{ for large enough } n, \tag{38}$$

then ϕ_n^{GCM} is LAUMP($\mathcal{S}$) against $\mathcal{L}_{\theta_n(h)}$ for $h \in (0,\infty) \times \mathcal{H}_g$, with

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n^{\mathrm{GCM}}(X,Y,Z)] = 1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)). \tag{39}$$

Let us discuss each of the four assumptions of Theorem 3:

- The assumption (35) is a set of regularity conditions on the null distributions $\mathscr{R}$. It is the same set of assumptions made by Shah and Peters (2020) to ensure Type-I error control of the GCM test over $\mathscr{R}$, including the assumption that the conditional means $\mu_{n,x}$ and $\mu_{n,y}$ are fit accurately enough (SP1) and fairly mild moment assumptions (SP2).
- The assumption (36) is a set of regularity conditions on the alternative distribution (29). These conditions are required for the semiparametric optimality theory to apply. These assumptions allow for GPLMs based on the normal distribution (assuming $\boldsymbol{X}$ has second moment) or any other exponential family (assuming $(\boldsymbol{X},\boldsymbol{Z})$ is compactly supported and the functions g are continuous).
- The assumption (37) states that the conditional expectation $\boldsymbol{Z} \mapsto \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\boldsymbol{Z}]$ must belong to the subspace $\mathcal{H}_g$. It guarantees that the "least favorable" value of the nonparametric component g is in the space $\mathcal{H}_g$, yielding the optimality of the GCM statistic.
- The assumption (38) connects the semiparametric alternative hypothesis to the conditional independence null hypothesis. In some sense it requires $\mathcal{L}_{\theta_0} \equiv \mathcal{L}_{(0,g_0)}$ (derived from the semiparametric alternative distribution (29)) to be an interior point of $\mathscr{R}$ (the conditional independence null) for each $g_0 \in \mathcal{S}$.

We give an example of when these assumptions hold in the next section.

5.2. *Example: Kernel ridge regression.* We illustrate Theorem 3 with a kernel ridge regression example, borrowed from Shah and Peters (2020, Section 4). Suppose the conditional expectations $\mu_x(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mu_y(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y}|\boldsymbol{Z}]$ satisfy $\mu_x, \mu_y \in \mathcal{H}_k$ for some reproducing kernel Hilbert space $(\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$ with reproducing kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In particular, we consider $\mathcal{H}_k \equiv W^{1,2}([0,1]) \subset L^2([0,1])$, i.e. the Sobolev space defined

$$W^{1,2}([0,1]) \equiv \big\{ f : [0,1] \to \mathbb{R} \mid f(0) = 0,\ f \text{ is absolutely continuous with } \dot{f} \in L^2([0,1]) \big\},$$

equipped with the inner product

$$\langle f, g \rangle_{W^{1,2}([0,1])} \equiv \int_0^1 \dot{f}(z)\dot{g}(z)\mathrm{d}z.$$

$W^{1,2}([0,1])$ is an RKHS with kernel $k(x,y) = \min\{x,y\}$ (Wainwright, 2019, Example 12.16). Consider the kernel ridge estimators

$$\begin{aligned} \widehat{\mu}_x &\equiv \underset{\mu_x \in W^{1,2}([0,1])}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n |X_i - \mu_x(Z_i)|^2 + \lambda \|\mu_x\|^2_{W^{1,2}([0,1])} \right\}; \\ \widehat{\mu}_y &\equiv \underset{\mu_y \in W^{1,2}([0,1])}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - \mu_y(Z_i)|^2 + \lambda \|\mu_y\|^2_{W^{1,2}([0,1])} \right\}, \end{aligned} \tag{40}$$

with λ tuned as described in Shah and Peters (2020, Section 4). Using Shah and Peters (2020, Theorem 11), the following result can be derived as a consequence of Theorem 3.

COROLLARY 4. *Fix $C > 0$, and consider the following regularity class $\mathscr{R} \subseteq \mathscr{L}^0$:*

$$\begin{aligned} \mathscr{R} \equiv \{ &\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \mathcal{L}(\boldsymbol{Z}) \times \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) \times \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) : \\ &\mathcal{L}(\boldsymbol{Z}) = \mathrm{Unif}([0,1]),\ \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = N(\mu_x(\boldsymbol{Z}), 1),\ \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\mu_y(\boldsymbol{Z}), 1), \\ &\mu_x, \mu_y \in B_{W^{1,2}}(0, C)\}, \end{aligned} \tag{41}$$

where we define the $W^{1,2}([0,1])$ ball

$$B_{W^{1,2}}(0, C) \equiv \{ f \in W^{1,2}([0,1]) : \|f\|_{W^{1,2}([0,1])} < C \}. \tag{42}$$

Now, fix $\mu_{0x}, \mu_{0y} \in B_{W^{1,2}}(0,C)$ and for each $h = (h_\beta, h_g) \in (0,\infty) \times W^{1,2}([0,1])$ consider the set of local alternatives $\mathcal{L}_{\theta_n(h)}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ given by

$$\begin{aligned} &\mathcal{L}_{\theta_n(h)}(\boldsymbol{Z}) \equiv \mathrm{Unif}([0,1]); \\ &\mathcal{L}_{\theta_n(h)}(\boldsymbol{X}|\boldsymbol{Z}) \equiv N(\mu_{0x}(\boldsymbol{Z}), 1); \\ &\mathcal{L}_{\theta_n(h)}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) \equiv N(\boldsymbol{X} h_\beta / \sqrt{n} + \mu_{0y}(\boldsymbol{Z}) + h_g(\boldsymbol{Z})/\sqrt{n}, 1). \end{aligned} \tag{43}$$

Then, the GCM test based on the kernel ridge estimators (40) is LAUMP($B_{W^{1,2}}(0,C)$) against alternatives $\mathcal{L}_{\theta_n(h)}$.

Hence, the GCM test based on kernel ridge regression does not just control Type-I error (Shah and Peters, 2020, Theorem 11); it is also optimal against local alternatives.

5.3. *Discussion of Theorem 3.* Theorem 3 states that the GCM test of Shah and Peters (2020) is the optimal test of conditional independence against a broad class of semiparametric GPLM alternatives, including linear and generalized linear models. To our knowledge, it is the first result at the intersection of conditional independence testing and semiparametric optimality, although Shah and Peters (2020) have already noted the connection between the

GCM test and nonparametric estimation of the expected conditional covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$. Our result complements another line of work on minimax optimality for conditional independence testing (Canonne et al., 2018; Neykov, Balakrishnan and Wasserman, 2021; Kim et al., 2022). In the related model-X context, few optimality results are available. Two existing works show optimality statements based on likelihood ratio statistics; one in the context of the CRT (Katsevich and Ramdas, 2022) and the other in the context of model-X knockoffs (Spector and Fithian, 2022).

Theorem 3 closely parallels results on estimation in semiparametric regression (Robinson, 1988; Bickel et al., 1993; Donald and Newey, 1994; Härdle, Liang and Gao, 2000; Robins and Rotnitzky, 2001; Van De Geer et al., 2014; Ning and Liu, 2017; Janková and Van De Geer, 2018; Chernozhukov et al., 2018). It follows from Bickel et al. (1993); Robins and Rotnitzky (2001) that the GCM statistic with the true conditional means μ_x and μ_y is the efficient score under the null hypothesis $\beta = 0$ in the context of GPLMs based on one-parameter exponential families with canonical link. Existing results on semiparametric optimality for hypothesis testing state that tests based on optimal estimators are themselves optimal (Choi, Hall and Schick, 1996; Van Der Vaart, 1998; Kosorok, 2008).

Despite the similarity between Theorem 3 and existing semiparametric optimality results, we emphasize that this theorem is a statement about optimality for conditional independence testing rather than for semiparametric testing. The semiparametric model (29) plays the role of the alternative distribution with respect to which power is evaluated, and need not hold under the null hypothesis. To bridge this gap, it suffices to find an open ball within the conditional independence null hypothesis containing the semiparametric null hypothesis (38). This allows us to reduce the conditional independence testing problem to a semiparametric testing problem, and therefore to leverage existing semiparametric optimality results (Appendix E in Niu et al. (2023)).

Note that Theorem 3 gives the power against local alternatives of the GCM test with μ_x and μ_y estimated in sample. This complements Shah and Peters (2020, Theorem 8), where these authors compute the power of the GCM test against non-local alternatives by resorting to sample splitting, which is not required to show Type-I error control for the GCM test. This sample splitting is necessary under non-local alternatives to avoid Donsker conditions; using either sample splitting or Donsker conditions is also standard practice in the semiparametric literature. By contrast, we avoid sample splitting by exploiting the special structure of the conditional independence null and contiguity arguments to compute limiting power under local alternatives.

While the Type-I error control results in Section 4 are stated in the high-dimensional setting, Theorem 3 is stated only for fixed-dimensional covariate vectors $\boldsymbol{Z}$. Indeed, semiparametric optimality theory is predominantly low-dimensional. A notable exception is the work of Janková and Van De Geer (2018), which provides a semiparametric theory of estimation in high dimensions. Extending this theory to hypothesis testing is nontrivial, and beyond the scope of the current work. Nevertheless, proving optimality statements for conditional independence testing in high dimensions is an interesting direction for future work. We note in passing that high-dimensional results for lasso-based estimators often assume exact sparsity of the coefficient vector, which poses a problem for condition (38) requiring the regularity class $\mathscr{R}$ to have interior points.

Finally, we note that Theorem 3 gives the optimality of the GCM statistic against alternative models for $\boldsymbol{Y}$ in which $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not interact. For alternatives where the conditional association between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is modified by $\boldsymbol{Z}$, the GCM test will no longer be optimal. Variants of the CRT (Zhong, Kuffner and Lahiri, 2021; Sesia and Sun, 2022), model-X knockoffs (Li et al., 2021), and the GCM test (Lundborg et al., 2022) are designed to improve power in the presence of effect modification are available, although their optimality properties are not described. Optimal tests developed specifically for detecting interaction effects between $\boldsymbol{X}$ and $\boldsymbol{Z}$ (rather than main effects) may be constructed based on Vansteelandt et al. (2008).

6. Finite-sample performance assessment. The results in the preceding sections are all asymptotic. In this section, we complement these results with a comprehensive simulation-based assessment of Type-I error and power in finite samples. Previous simulation-based assessments of the Type-I error of MX methods have come to differing conclusions: Sesia, Sabatti and Candès (2019); Romano, Sesia and Candès (2019); Sesia et al. (2020); Liu et al. (2022) found broad robustness to misspecification of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ while Li and Liu (2022) found such misspecifications to cause marked Type-I error inflation. We show that differences in the level of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ implied by the simulation design explain these discrepancies, and then use this insight to inform our own simulation design in Section 6.2. Then, we present the results of our numerical simulations in Section 6.3. Numerical simulation results and instructions to reproduce them are available at `https://github.com/Katsevich-Lab/symcrt-manuscript`.

6.1. *Revisiting prior simulations of robustness.* The question of robustness of MX methods to the misspecification of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ has been investigated starting from the paper in which the model-X framework was originally proposed (Candès et al., 2018). In this paper, the joint distribution $\mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Z})$ was estimated in sample via the graphical lasso, which is similar to estimating the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ via the ordinary lasso. These authors found that

> "Although the graphical Lasso is well suited for this problem since the covariates have a sparse precision matrix, its covariance estimate is still off by nearly 50%, and yet surprisingly the resulting power and FDR are nearly indistinguishable from when the exact covariance is used...the nominal level of 10% FDR is never violated, even for covariance estimates very far from the truth."

Similar conclusions have been drawn from numerical simulations in subsequent papers as well (Sesia, Sabatti and Candès, 2019; Romano, Sesia and Candès, 2019; Sesia et al., 2020; Liu et al., 2022), the latter studying the dCRT specifically. On the other hand, the numerical simulations of Li and Liu (2022) show that the dCRT can suffer significant Type-I error inflation when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is inaccurately fit. These authors state that "for model-X inference, the dependence of $\boldsymbol{X}$ on $\boldsymbol{Z}$ is not adequately characterized and adjusted [for] due to the shrinkage bias of lasso."

To resolve this apparent contradiction, we consider a common data-generating model used in MX literature:

$$\mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Z}) = N(0, \Sigma), \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}\theta + \boldsymbol{Z}^T\beta, \sigma_y^2). \tag{44}$$

Often, $(\boldsymbol{X}, \boldsymbol{Z})$ are assumed to have a spatial structure (motivated by the GWAS application), with $\Sigma = \Sigma(\rho) \in \mathbb{R}^{(1+p)\times(1+p)}$ taken to be the AR(1) covariance matrix with autocorrelation parameter $\rho \in (-1, 1)$. This covariance matrix roughly approximates linkage disequilibrium structure among genotypes, where correlations among variables are local with respect to the spatial structure. Conditional independence under this model (44) reduces to $H_0 : \theta = 0$. Furthermore, the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ implied by the normal joint distribution is that of a linear model:

$$\text{Under } H_0, \quad \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\gamma, \sigma_x^2), \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, \sigma_y^2). \tag{45}$$

In the context of this model, the conditional independence testing problem is nontrivial to the extent that $\boldsymbol{Z}$ induces marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ even in the absence of conditional association. In a causal inference context, this spurious marginal association would be called a confounding effect of $\boldsymbol{Z}$. This marginal association can be small or large, depending on the correlation structure of $\boldsymbol{Z}$ and the extent to which the supports of β and γ overlap. Properly adjusting for $\boldsymbol{Z}$ is important to the extent that $\boldsymbol{Z}$ induces marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

We claim that the simulation studies in much of the original MX literature had relatively low levels of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$, whereas the simulation studies in Li and Liu (2022) were done in a regime with much more marginal association. To illustrate this point, we quantify the level of marginal association in a given problem setup as the Type-I error of the GCM test with intercept-only models for $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$. This test is essentially a Pearson test of (marginal) independence between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and ignores the variables $\boldsymbol{Z}$ altogether. We compute this Type-I error for the data-generating models used to assess robustness by Candès et al. (2018); Liu et al. (2022); Li and Liu (2022) (Appendix F.1 in Niu et al. (2023)). The former two papers are framed in the variable selection context, where several explanatory variables $\boldsymbol{W}_j$ are considered, and the hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{\text{-}j}$ is tested for each j. Therefore, $\boldsymbol{X} \equiv \boldsymbol{W}_j$ for each j. On the other hand, Li and Liu (2022) considered a conditional independence testing framework, where $\boldsymbol{X}$ was a single variable of interest.

For the data-generating models used by Candès et al. (2018); Liu et al. (2022), we evaluate the Type-I error of the marginal GCM test for each hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{\text{-}j}$, plotting these as a function of j (Figure 2, top row). We superimpose onto these plots a blue horizontal line indicating the Type-I error of the marginal GCM test (fitting the intercept only model) for the data-generating model used by Li and Liu (2022) (equal to 0.99, suggesting strong marginal association), and a red dashed horizontal line indicating the nominal level of this marginal test (equal to 0.05). The green ticks indicate the locations of the non-null variables. As expected for a setting where variable correlation is local, we see that Type-I error is inflated for null variables near the signal variables. The extent of this inflation depends on the autocorrelation parameter (set at 0.3 by Candès et al. (2018) and 0.5 by Liu et al. (2022)) and the locations of the signal variables. Most null variables, however, are not near signal variables, and therefore the marginal GCM test shows no inflation. This is reflected by the histograms of the Type-I error inflations (Figure 2, bottom row). The median Type-I error of the marginal GCM test is near the nominal level of 0.05 in all three of the simulation setups from Candès et al. (2018); Liu et al. (2022).

6.2. *Simulation design.*

Data-generating model. As discussed in the previous section, appropriately setting the marginal correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a given data-generating model is crucial to properly evaluate the impact of inaccurate estimation of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ on the Type-I error control of a model-X method. Keeping this in mind, we propose the following data-generating model:

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, \Sigma(\rho)),\ \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1),\ \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}\theta + \boldsymbol{Z}^T\beta, 1). \tag{46}$$

We set the first s coefficients of β to be equal to ν and the rest to zero. Therefore, the entire data-generating process is parameterized by the six parameters $(n, p, s, \rho, \theta, \nu)$ (Table 1). For both null and alternative simulations, we vary each of the first four across five values each, setting the remaining three to the default value indicated in bold. The fifth parameter θ controls the signal strength and the sixth parameter ν controls the extent of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For the null simulation, we set $\theta \equiv 0$, and for each setting of (n, p, s, ρ), we choose five values of ν equally spaced between 0 (no marginal association) and $\nu_{\max}$ (computed so that the marginal GCM method has Type-I error 0.99). Note that $\nu_{\max}$ depends on the parameters (n, p, s, ρ), so not exactly the same values of ν were used across settings of these four parameters. For the alternative simulation, we kept ν fixed at $\nu_{\max}/2$ while for each setting of (n, p, s, ρ), we choose five values of θ equally spaced between 0 (no signal) and $\theta_{\max}$ (computed so that the GCM method with oracle settings of $\widehat{\mu}_{n,x}$ and $\widehat{\mu}_{n,y}$ has power 0.99). Finally, we complement the linear regression data-generating model (46) with an analogous one based on logistic regression.

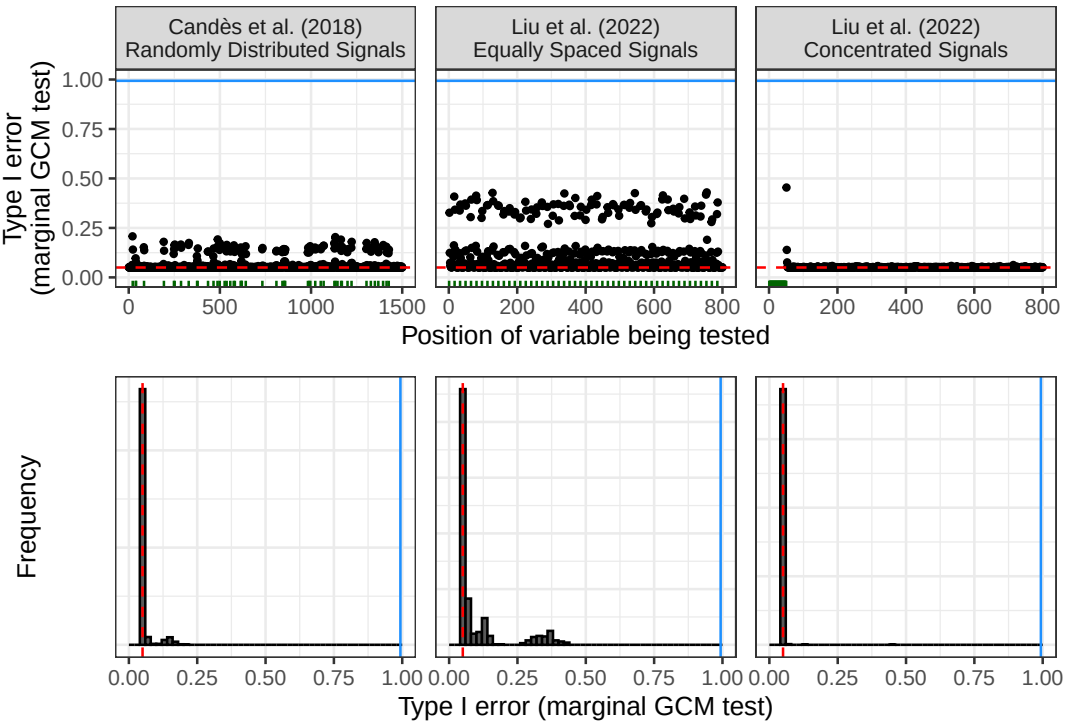

Fig 2: Comparing the marginal associations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the robustness simulations of Candès et al. (2018); Liu et al. (2022); Li and Liu (2022) (Appendix F.1 in Niu et al. (2023)). Top: Type-I error of the marginal GCM test as a function of the position of null variables with respect to the non-null variables (represented as green ticks). Bottom: Histograms of the Type-I error across null variables. The solid blue line indicates the Type-I error of the marginal GCM test for the robustness simulation of Li and Liu (2022), and the dashed red line the nominal Type-I error level of the marginal GCM test (0.05).

n	p	s	ρ
100	100	**5**	0
200	200	10	0.2
400	**400**	20	**0.4**
800	800	40	0.6
1600	1600	80	0.8

θ (null)	ν (null)
0	0
0	$\nu_{\max}/4$
0	$\nu_{\max}/2$
0	$3\nu_{\max}/4$
0	$\nu_{\max}$

θ (alt)	ν (alt)
0	$\nu_{\max}/2$
$\theta_{\max}/4$	$\nu_{\max}/2$
$\theta_{\max}/2$	$\nu_{\max}/2$
$3\theta_{\max}/4$	$\nu_{\max}/2$
$\theta_{\max}$	$\nu_{\max}/2$

TABLE 1

The values of the sample size n, covariate dimension p, sparsity s, autocorrelation of covariates ρ, signal strength θ, and marginal association strength ν used for the simulation study. Each of the parameters n, p, s, ρ was varied among the values in the first table while keeping the other three at their default values, indicated in bold. For example, $p = 400, s = 5, \rho = 0.4$ were kept fixed while varying $n \in \{100, 200, 400, 800, 1600\}$. The second and third tables denote the values of (θ, ν) used for the null and alternative simulations. Each combination of (n, p, s, ρ) was paired with each of the five values of (θ, ν) displayed for null and alternative simulations.

Methodologies compared. In Section 4, we found that the GCM test and the $\widehat{\text{dCRT}}$ are equivalent when applied with the same estimation methods for $\mu_{n,x}$ and $\mu_{n,y}$. Using this equivalence, we also showed that the $\widehat{\text{dCRT}}$ is robust to errors in $\widehat{\mu}_{n,x}$ if they are compensated for by accurate estimates $\widehat{\mu}_{n,y}$. In our simulation to assess Type-I error, we wish to probe the finite-sample Type-I error control of the GCM and the $\widehat{\text{dCRT}}$. We apply both of these

methods with the lasso to estimate $\mu_{n,x}$ and $\mu_{n,y}$, as this is the most common choice in the MX literature.

In addition to the GCM test and the $\widehat{\text{dCRT}}$, we apply the Maxway CRT (Li and Liu, 2022), designed specifically to improve the Type-I error control of the dCRT in the context when $\mu_{n,x}$ must be estimated. The Maxway CRT is inherently a semi-supervised method, assuming the existence of an auxiliary unlabeled dataset containing observations of $\boldsymbol{X}$ and $\boldsymbol{Z}$ but not of $\boldsymbol{Y}$. The methodology (specifically, "Maxway$_{\text{in}}$ example 1") proceeds—roughly—by fitting $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ on the unlabeled data via the post-lasso (i.e. selecting active variables via the lasso and then refitting via ordinary least squares, Belloni and Chernozhukov (2013)), fitting $\widehat{\mu}_{ny}(\boldsymbol{Z})$ on the labeled data via post-lasso, and then applying dCRT on the labeled data based on these two models.

Since the primary focus of this paper is the setting when no auxiliary unlabeled data are available, we implement the Maxway CRT by randomly splitting the data into two equal pieces, using the first as the unlabeled data (in particular, ignoring the response data) and the second as the labeled data. This strategy is consistent with the real data analysis in Li and Liu (2022, Section 6). We also consider a bona-fide semi-supervised setup, in order to compare the GCM test and $\widehat{\text{dCRT}}$ to the Maxway CRT in the setting originally considered by Li and Liu (2022). However, in the semi-supervised setting we use all of the available data on $(\boldsymbol{X}, \boldsymbol{Z})$ (i.e. both unlabeled and labeled data) to fit $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. By contrast, Li and Liu (2022) used only the unlabeled data to learn $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ in their implementation of the $\widehat{\text{dCRT}}$ for semi-supervised data.

Finally, we noted in Section 4 that the $\widehat{\text{dCRT}}$ already has a built-in doubly robust property. Therefore, we conjectured that the Type-I error inflation observed in the simulations of Li and Liu (2022) is attributable to poor estimation of $\mu_n(\boldsymbol{X}|\boldsymbol{Z})$ and/or $\mu_n(\boldsymbol{Y}|\boldsymbol{Z})$ and that the $\widehat{\text{dCRT}}$ can achieve Type-I error control if used in conjunction with better estimators of these conditional means. Taking inspiration from Li and Liu (2022), we also considered versions of the $\widehat{\text{dCRT}}$ and the GCM test based on the post-lasso in addition to those based on the usual lasso. In summary, we compared five methods: lasso and post-lasso based GCM, lasso and post-lasso based $\widehat{\text{dCRT}}$, and Maxway CRT (Table 2). As a point of reference for the null simulation, we also included the GCM test with intercept-only models for $\mu_{n,x}$ and $\mu_{n,y}$; the Type-I error of this test quantifies the degree of marginal association in the data-generating model (Section 6.1). As a point of reference for the alternative simulation, we also included the GCM test with $\mu_{n,x}$ and $\mu_{n,y}$ set to their ground truth values; the power of this test is the maximum power achievable by any test and therefore quantifies the signal strength in the data-generating model.

Evaluation of power in the presence of Type-I error inflation. The methodologies compared control Type-I error to differing extents across the variety of simulation parameters in Table 1. This makes it challenging to compare power across methods, since some control Type-I error while others do not. To address this challenge, we chose to compare the power of the *test statistics* underlying the methods, each under oracle calibration to ensure Type-I error control. Given the composite null, exact oracle calibration is computationally intractable. Therefore, we instead calibrated each test with respect to the point null given by

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, \Sigma(\rho)),\ \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1),\ \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]^T\theta + \boldsymbol{Z}^T\beta, 1).$$

This is the "closest" point in the null to the alternative (46) under consideration; therefore ensuring Type-I error control at this point null should be a decent proxy for ensuring Type-I error control over the whole null. To calibrate two-sided tests with respect to this point null, we generate samples of a test statistic from the null and then define lower and upper critical values as the 2.5% and 97.5% quantiles of this distribution. Using potentially asymmetric

Method name	Estimating $\mu_{n,x}$	Data for $\widehat{\mu}_{n,x}$	Estimating $\mu_{n,y}$	Data for $\widehat{\mu}_{n,y}$
GCM (LASSO)	lasso	all	lasso	all/labeled
$\widehat{\text{dCRT}}$ (LASSO)	lasso	all	lasso	all/labeled
GCM (PLASSO)	post-lasso	all	post-lasso	all/labeled
$\widehat{\text{dCRT}}$ (PLASSO)	post-lasso	all	post-lasso	all/labeled
Maxway CRT	post-lasso	unlabeled	post-lasso	labeled
GCM (marginal)	intercept-only	all	intercept-only	all/labeled
GCM (oracle)	ground truth	–	ground truth	–

TABLE 2

The five methodologies compared, how they estimate $\mu_{n,x}$ and $\mu_{n,y}$, and what data they use for each in the context of semi-supervised or fully supervised data. Note that in the fully supervised case, data is split in half to form "unlabeled" and labeled sets for Maxway CRT. In this case, the $\widehat{\text{dCRT}}$ and GCM tests still use all of the data available for estimating $\mu_{n,x}$ and $\mu_{n,y}$. Two additional tests were used for reference purposes: the GCM test with intercept-only models for $\mu_{n,x}$ and $\mu_{n,y}$ and the GCM test with $\mu_{n,x}$ and $\mu_{n,y}$ set to their ground truth values.

lower and upper critical values is necessary, as the null distribution may not be symmetric and centered at zero (Liu et al., 2022).

6.3. *Simulation results.* We conducted simulations for Gaussian and binary models for the response $\boldsymbol{Y}$, each within the supervised and semi-supervised settings. We present the Type-I error and power for Gaussian responses in the supervised setting in Figures 3 and 4, respectively, while deferring the other cases to Appendix F.3 in Niu et al. (2023). Note also that for the sake of brevity Figures 3 and 4 only present three out of the five values for the four parameters n, p, s, ρ; the complete results are presented in Appendix F.3 in Niu et al. (2023).

Next we list the main conclusions regarding Type-I error based on the results including figures in main text (Figure 3 (Gaussian supervised)), and figures in the supplementary material(Figure 4 (Gaussian semi-supervised), Figure 6 (binary supervised), and Figure 8 (binary semi-supervised) in Niu et al. (2023)):

- As one would expect, across all simulation settings, all methods have poorer Type-I error control as sample size n decreases, dimension p increases, number of nonzero coefficients s increases, autocorrelation ρ increases, or marginal association strength ν increases.
- For Gaussian responses, the $\widehat{\text{dCRT}}$ and GCM methods based on the same test statistics have very similar Type-I error control, echoing the asymptotic equivalence of the two methods (Theorem 2). For binary responses, the lasso-based $\widehat{\text{dCRT}}$ has somewhat lower Type-I error than the lasso-based GCM test (Figure 6 in Niu et al. (2023)). The discreteness of binary responses likely slows down the convergence to normality of the GCM statistic, rendering the resampling-based null distribution of the $\widehat{\text{dCRT}}$ a better approximation to the null distribution. We explore this phenomenon further in Appendix G.2 in Niu et al. (2023).
- Across all simulation settings, the $\widehat{\text{dCRT}}$ and GCM methods based on the post-lasso have dramatically better Type-I error control than their lasso-based counterparts. This is because the post-lasso tends to more fully regress the confounders $\boldsymbol{Z}$ out of the response $\boldsymbol{Y}$; see also Appendix F.2 in Niu et al. (2023).
- Across all simulation settings, Maxway CRT has better Type-I error control than the lasso-based $\widehat{\text{dCRT}}$ (in line with the results of Li and Liu (2022)), but worse Type-I error control than the post-lasso-based $\widehat{\text{dCRT}}$. The latter is likely due to the fact that Maxway CRT uses only half of the available data on $(\boldsymbol{X}, \boldsymbol{Z})$ to fit $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$, and therefore does not adjust for $\boldsymbol{Z}$ as accurately.

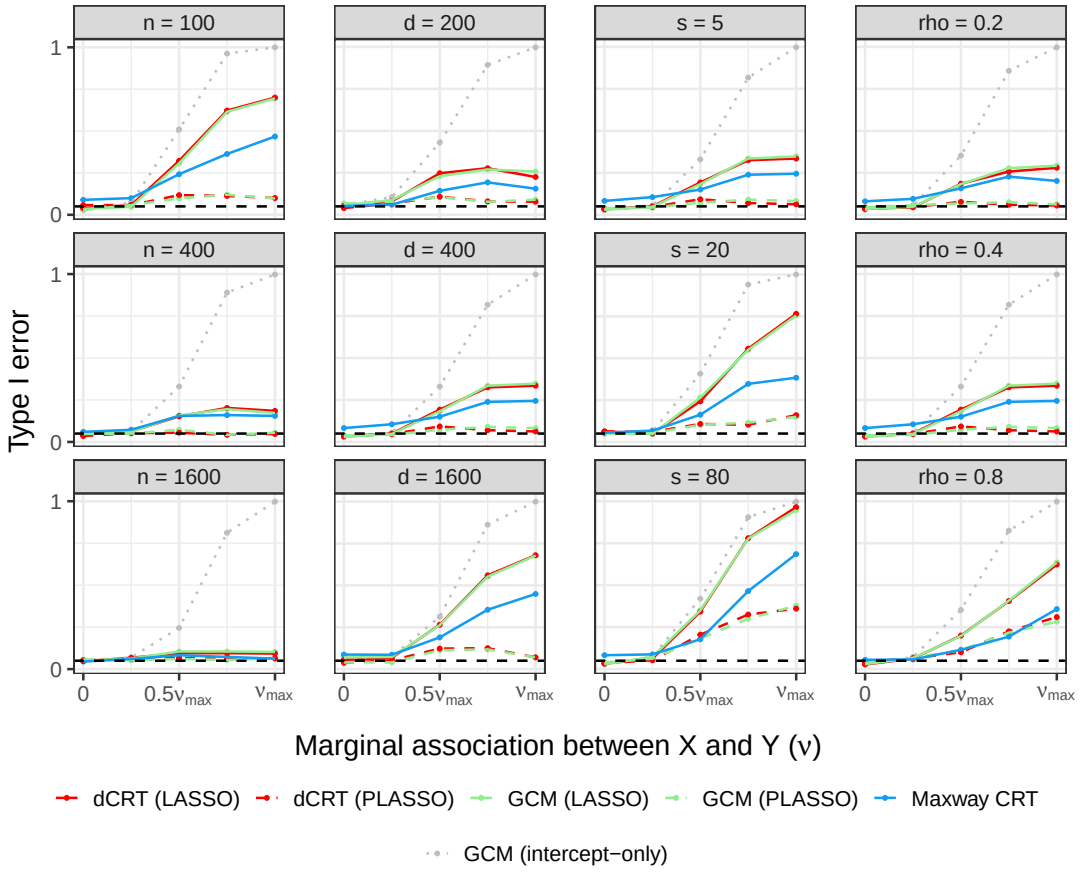

Fig 3: Type I error control for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the marginal association strength ν in each subplot. Each point is the average of 400 Monte Carlo replicates. All the standard errors are less than 0.026.

Next, we list the main conclusions regarding power based on the results including figures in the main paper(Figure 4 (Gaussian supervised)), and figures in supplementary material(Figure 5 (Gaussian semi-supervised), Figure 7 (binary supervised), and Figure 9 (binary semi-supervised) in Niu et al. (2023)):

- Across all simulation settings, GCM-based methods have somewhat higher power than their $\widehat{\text{dCRT}}$-based methods. This may have to do with the stabilizing effect of the GCM normalization, compared to the unnormalized $\widehat{\text{dCRT}}$ statistic. The difference between the two tends to vanish as sample size grows, reflecting the asymptotic equivalence of the two methods (Corollary 2).
- Across all simulation settings, the $\widehat{\text{dCRT}}$ and GCM methods based on the lasso have lower power than their post-lasso-based counterparts. This is because the post-lasso introduces more variance into the estimation of $\mu_{n,y}$; see also Appendix F.2 in Niu et al. (2023).
- Across Gaussian and binary supervised simulation settings (Figures 3 and 7 in Niu et al. (2023)), Maxway CRT has the lowest power among all methods compared. The reason for this is that Maxway CRT relies on data splitting and therefore has half the effective sample size of the other methods. On the other hand, for semi-supervised settings (Figures 5 and 9 in Niu et al. (2023)), Maxway CRT has power comparable to or better than those of the

post-lasso-based methods, but still worse than the lasso-based methods. This is due to the additional variance introduced by the refitting step in the post-lasso.

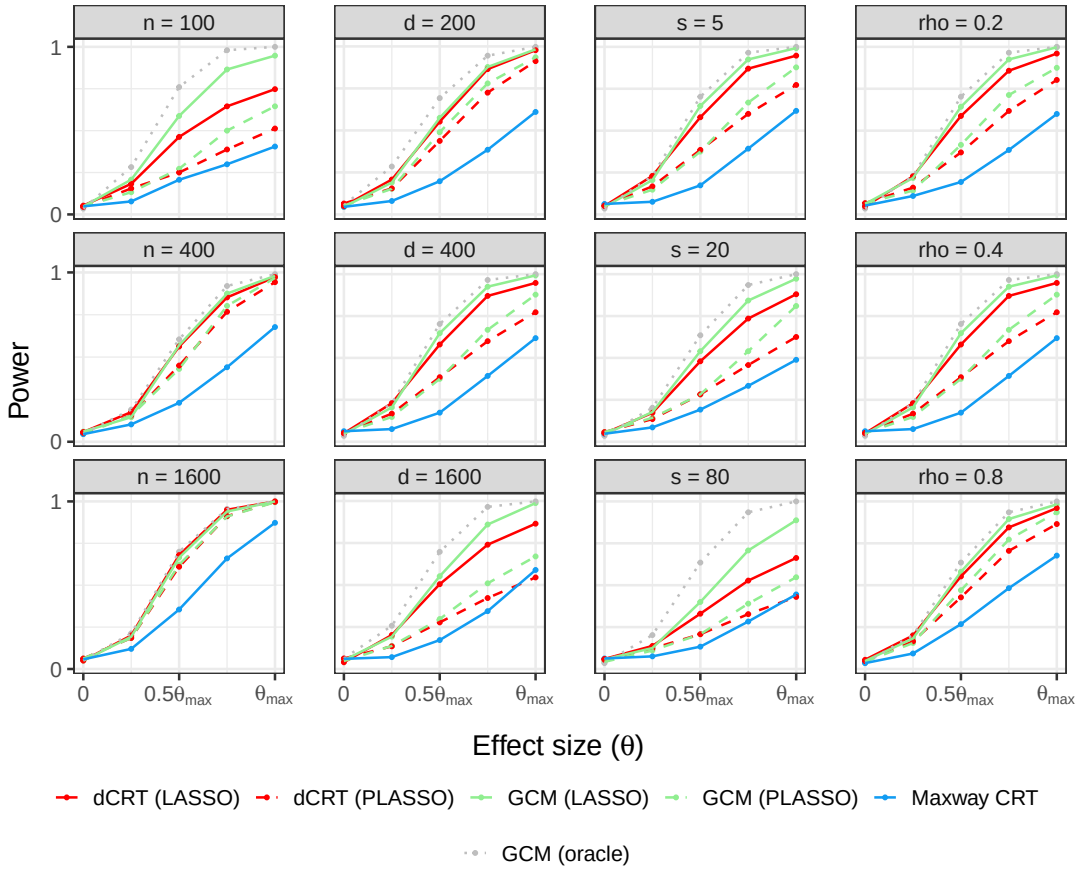

Fig 4: Power for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the signal strength θ in each subplot. Each point is the average of 400 Monte Carlo replicates. All the standard errors are less than 0.026.

In summary, the methods with the best Type-I error control across all simulation settings are the $\widehat{\text{dCRT}}$ and the GCM test based on the post-lasso, although this improved robustness does come with a cost in terms of power when compared to the lasso-based methods. We investigate the associated trade-off in Appendix F.2 in Niu et al. (2023).

7. Conclusion. We conclude by summarizing our main findings and highlighting directions for future work.

Model-X inference with $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ fit in sample can be doubly robust. Model-X inference (Candès et al., 2018) is presented as a mode of inference where the assumptions are transferred entirely from $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z})$ to $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$; no restrictions are made on the former law (or the test statistic used, at least in the context of the CRT), while the latter law is assumed exactly known. In practice, however, the law $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is often fit in sample. In the context of the dCRT, we show that Type-I error control cannot be guaranteed without restrictions on $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z})$ or the test statistic used (Section 3). On the other hand, test statistics based on

decent estimates of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ can compensate for errors in the estimation of $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ and restore Type-I error control (Corollary 3), a double robustness phenomenon. This result brings model-X inference more in line with double regression inferential methodologies: The conditional mean $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ is estimated in the context of in-sample approximation to the "model for X," and the conditional mean $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ is estimated when computing the model-X test statistic. Relatedly, a double robustness property was noted for conditional model-X knockoffs (Huang and Janson, 2020). A doubly robust version of the dCRT has also been recently proposed (the Maxway CRT; Li and Liu (2022)), although we argue that the original dCRT is itself doubly robust.

The GCM test has broadly similar Type-I error and power as the dCRT for large enough sample sizes, but requires no resampling. When fitting $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ in sample, the dCRT is essentially a double regression methodology. This prompts a comparison to the GCM test (Shah and Peters, 2020), another conditional independence test based on double regression. We established that the two tests are asymptotically equivalent under the null (Theorem 2) and under arbitrary local alternatives (Corollary 2). This suggests that the dCRT and the GCM test—when applied with the same estimators for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$—should have similar Type-I error control and power. Our numerical simulations (Section 6) largely confirm this behavior in finite samples. An exception to this conclusion is the case when small samples or discreteness in the data slows down the convergence of the GCM null distribution to normality. In such cases, we observed that the $\widehat{\text{dCRT}}$ can in fact have better Type-I error control than the GCM based on the same estimators (Figures 6 and 10 in Niu et al. (2023)), thanks to a better approximation to the null distribution in finite samples. Nevertheless, the broad similarity between the performances of the GCM test and the dCRT and the fact that the former test requires no resampling suggest that the GCM test may be preferable to the dCRT in practical problems with relatively large sample sizes.

The post-lasso yields much better Type-I error control than the lasso. Double robustness results for the GCM test and the dCRT apply only insofar as the estimation methods used in conjunction with these tests are accurate enough (SP1). The default estimation method for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ in many model-X applications is the lasso. As was demonstrated by Li and Liu (2022), the shrinkage bias of the lasso leads to inadequate adjustment of $\boldsymbol{X}$ and $\boldsymbol{Y}$ for $\boldsymbol{Z}$, which in turn leads to inflated Type-I error. The same authors proposed the Maxway CRT, an extension of the dCRT involving the identification of coordinates of $\boldsymbol{Z}$ impacting $\boldsymbol{X}$ and $\boldsymbol{Y}$ via the lasso followed by least squares refitting. Inspired by this work, we applied the original dCRT with post-lasso estimates for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$. We found vastly improved Type-I error control (Figure 2 in Niu et al. (2023)), compared not just to the lasso-based dCRT but also to the Maxway CRT itself. The decreased bias of the post-lasso helps adjust for $\boldsymbol{Z}$ more fully, although we found that the extra variance incurred by refitting does come at a cost in power. Nevertheless, our results suggest that applying the post-lasso in conjunction with model-X methodologies can lead to significant improvements in robustness.

The GCM test is the optimal conditional independence test against alternatives without interactions between $\boldsymbol{X}$ and $\boldsymbol{Z}$. It is widely known in the semiparametric literature that the GCM test is the efficient score test for (generalized) partially linear models. The connection between the GCM test and semiparametric theory was noted briefly by Shah and Peters (2020), though not explored in depth; presumably because the GCM test is a conditional independence test rather than a test of a parameter in a semiparametric model. Nevertheless, we find that if the semiparametric *null* hypothesis can be embedded within the conditional independence null hypothesis (38), semiparametric optimality theory can be carried over fairly directly to conditional independence testing to establish optimality against semiparametric

alternative distributions (Theorem 3). Thanks to this connection, we find that the GCM test has optimal asymptotic power among conditional independence tests against local generalized partially linear model alternatives (29). On the other hand, we leave open the question of optimality against alternatives where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are allowed to interact. We also leave open whether our optimality result can be extended to the high-dimensional regime.

Future work: The proportional regime, other test statistics,. and the variable selection problem

Our results about the equivalence between the GCM test and the dCRT, and the double robustness of the latter, require estimates of $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ that are individually consistent and whose rates of convergence are sufficiently fast (SP1). In the case of sparse linear models, we can get such rates if $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ depend on at most $s = o(\sqrt{n}/\log(p))$ of the coordinates of $\boldsymbol{Z}$. Such assumptions are common in other lines of work on high-dimensional / semiparametric / doubly robust inference, including the debiased lasso (Van De Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Ning and Liu, 2017; Janková and Van De Geer, 2018) and doubly robust causal inference (Belloni, Chernozhukov and Hansen, 2014; Chernozhukov et al., 2018). On the other hand, consistent estimates are typically not available in the regime when n, p, and s grow proportionally (Bayati and Montanari, 2011), causing a failure in traditional debiased estimates (Celentano and Montanari, 2021). An additional limitation of the current work is that we prove double robustness of the CRT for only one test statistic, namely the dCRT statistic. A natural question to ask is whether this property is enjoyed by a broader class of test statistics. This may be accomplished by proving equivalence of the CRT based on other doubly robust test statistics to the corresponding asymptotic tests. However, this would likely entail deriving the limiting CRT resampling distribution (analogously to Section 2), which may be harder for test statistics whose dependence on X is more complex than that of the dCRT statistic. Finally, we did not directly consider the variable selection problem or the MX knockoffs procedure in the current work. We conjecture that MX knockoffs also enjoys some notion of double robustness; indirect evidence for this was presented recently (Fan, Gao and Lv, 2023). It would also be interesting to explore whether MX knockoffs enjoys any optimality properties as a variable selection procedure, though this is a complex question because its power is a function of not just the test statistic choice but also of the knockoff filter multiple testing procedure.

Acknowledgments. We acknowledge help from Timothy Barry with our simulation studies and the underlying computational infrastructure, including his `simulatr` R package and Nextflow pipeline. We acknowledge dedicated support from the staff at the Wharton High Performance Computing Cluster. We acknowledge Lucas Janson for providing details about the simulation setting in Candès et al. (2018). We acknowledge Eric Tchetgen Tchetgen for helpful discussions on hypothesis testing in the semiparametric models. Finally, we acknowledge two referees and an associate editor for their insightful comments and suggestions, which helped improve this work.

Funding. ZN was partially supported by the grant "Statistical Software for Single Cell CRISPR Screens" awarded to EK by Analytics at Wharton. OD was partially supported by FWO grant 1222522N and NIH grant AG065276. EK was partially supported by NSF DMS-2113072 and NSF DMS-2310654.

SUPPLEMENTARY MATERIAL

Supplement to: "Reconciling model-X and doubly robust approaches to conditional independence testing"

This supplement includes all the proofs of the results in the main paper and additional simulation results.

REFERENCES

AUFIERO, M. and JANSON, L. (2022). Surrogate-based global sensitivity analysis with statistical guarantees via floodgate. *arXiv*.

BARBER, R. F., CANDÈS, E. J. and SAMWORTH, R. J. (2020). Robust inference with knockoffs. *Annals of Statistics,*.

BARBER, R. F. and JANSON, L. (2022). Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *Annals of Statistics* **50** 2514–2544.

BARRY, T., WANG, X., MORRIS, J. A., ROEDER, K. and KATSEVICH, E. (2021). SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biology*.

BATES, S., SESIA, M., SABATTI, C. and CANDES, E. (2020). Causal Inference in Genetic Trio Studies. *Proceedings of the National Academy of Sciences* **117** 24117–24126.

BAYATI, M. and MONTANARI, A. (2011). The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory* **58** 1997–2017. https://doi.org/10.1109/TIT.2011.2174612

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. https://doi.org/10.3150/11-BEJ410

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

BERRETT, T. B., WANG, Y., FOYGEL BARBER, R. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **82** 175–197.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. A. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 551–577.

CANONNE, C. L., DIAKONIKOLAS, I., KANE, D. M. and STEWART, A. (2018). Testing conditional independence of discrete distributions. *2018 Information Theory and Applications Workshop, ITA 2018* 735–748. https://doi.org/10.1109/ITA.2018.8503255

CELENTANO, M. and MONTANARI, A. (2021). CAD: Debiasing the Lasso with inaccurate covariate model. *arXiv*.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* **21** C1–C68. https://doi.org/10.1111/ectj.12097

CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H., NEWEY, W. K. and ROBINS, J. M. (2022). Locally Robust Semiparametric Estimation. *Econometrica* **90** 1501–1535. https://doi.org/10.3982/ecta16294

CHOI, S., HALL, W. J. and SCHICK, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Annals of Statistics* **24** 841–861. https://doi.org/10.1214/aos/1032894469

DONALD, S. G. and NEWEY, W. K. (1994). Series estimation of semilinear models. https://doi.org/10.1006/jmva.1994.1032

DUKES, O. and VANSTEELANDT, S. (2020). How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical Methods in Medical Research* **29** 677–694. https://doi.org/10.1177/0962280219862005

FAN, Y., GAO, L. and LV, J. (2023). ARK: Robust Knockoffs Inference with Coupling. *arXiv*.

FAN, Y., LV, J., SHARIFVAGHEFI, M. and UEMATSU, Y. (2020a). IPAD: Stable Interpretable Forecasting with Knockoffs Inference. *Journal of the American Statistical Association*. https://doi.org/10.2139/ssrn.3245137

FAN, Y., DEMIRKAYA, E., LI, G. and LV, J. (2020b). RANK: Large-Scale Inference With Graphical Nonlinear Knockoffs. *Journal of the American Statistical Association* **115** 362–379. https://doi.org/10.1080/01621459.2018.1546589

HAM, D. W., IMAI, K. and JANSON, L. (2022). Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis. *arXiv*.

HÄRDLE, W., LIANG, H. and GAO, J. (2000). *Partially linear models*. Springer Science & Business Media.

HENMI, M. and EGUCHI, S. (2004). A Paradox concerning Nuisance Parameters and Projected Estimating Functions. *Biometrika* **91** 929–941.

HUANG, D. and JANSON, L. (2020). Relaxing the Assumptions of Knockoffs by Conditioning. *Annals of Statistics* **48** 3021–3042.

JANKOVÁ, J. and VAN DE GEER, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *Annals of Statistics* **46** 2336–2359. https://doi.org/10.1214/17-AOS1622

JAVANMARD, A. and MONTANARI, A. (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research* **15** 2869–2909.

KATSEVICH, E. and RAMDAS, A. (2022). On the power of conditional independence testing under model-X. *Electronic Journal of Statistics* **16** 6348–6394.

KIM, I., NEYKOV, M., BALAKRISHNAN, S. and WASSERMAN, L. (2022). Local permutation tests for conditional independence. *Annals of Statistics* **50** 3388–3414.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

LI, S. and LIU, M. (2022). Maxway CRT: Improving the Robustness of Model-X Inference. *arXiv*.

LI, S., SESIA, M., ROMANO, Y., CANDÈS, E. and SABATTI, C. (2021). Searching for consistent associations with a multi-environment knockoff filter. *Biometrika*.

LIU, J. and RIGOLLET, P. (2019). Power analysis of knockoff filters for correlated designs. In *33rd Conference on Neural Information Processing Systems*.

LIU, M., KATSEVICH, E., JANSON, L. and RAMDAS, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109** 277–293. https://doi.org/10.1093/biomet/asab039

LUNDBORG, A. R., KIM, I., SHAH, R. D. and SAMWORTH, R. J. (2022). The Projected Covariance Measure for assumption-lean variable significance testing. *arXiv*.

NEYKOV, M., BALAKRISHNAN, S. and WASSERMAN, L. (2021). Minimax optimal conditional independence testing. *Annals of Statistics* **49** 2151–2177. https://doi.org/10.1214/20-AOS2030

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics* **45** 158–195. https://doi.org/10.1214/16-AOS1448

NIU, Z., CHARABORTY, A., DUKES, O. and KATSEVICH, E. (2023). Supplement to "Reconciling model-X and doubly robust approaches to conditional independence testing".

PEARL, J. (2009). *Causality*. Cambridge University Press.

ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders. *Biometrics* **48** 479–495.

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11** 920–936.

ROBINSON, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica* **56** 931–954.

ROMANO, Y., SESIA, M. and CANDÈS, E. (2019). Deep Knockoffs. *Journal of the American Statistical Association* **0** 1–27. https://doi.org/10.1080/01621459.2019.1660174

SESIA, M., SABATTI, C. and CANDÈS, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. https://doi.org/10.1093/biomet/asy033

SESIA, M. and SUN, T. (2022). Individualized conditional independence testing under model-X with heterogeneous samples and interactions. *arXiv*.

SESIA, M., KATSEVICH, E., BATES, S., CANDÈS, E. and SABATTI, C. (2020). Multi-resolution localization of causal variants across the genome. *Nature Communications* **11** 1093.

SESIA, M., BATES, S., CANDÈS, E., MARCHINI, J. and SABATTI, C. (2021). False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences of the United States of America* **118** 1–12. https://doi.org/10.1073/pnas.2105841118

SHAH, R. D. and PETERS, J. (2020). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *Annals of Statistics* **48** 1514–1538.

SMUCLER, E., ROTNITZKY, A. and ROBINS, J. M. (2019). A unifying approach for doubly-robust L1 regularized estimation of causal contrasts. *arXiv*.

SPECTOR, A. and FITHIAN, W. (2022). Asymptotically Optimal Knockoff Statistics via the Masked Likelihood Ratio. *arXiv*.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* **42** 1166–1202. https://doi.org/10.1214/14-AOS1221

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

VANSTEELANDT, S., VANDERWEELE, T. J., TCHETGEN, E. J. and ROBINS, J. M. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association* **103** 1693–1704. https://doi.org/10.1198/016214508000001084

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. https://doi.org/10.1017/9781108627771

WANG, W. and JANSON, L. (2022). A high-dimensional Power Analysis of the Conditional Randomization Test and Knockoffs. *Biometrika* **109** 631–645.

WEINSTEIN, A., BARBER, R. and CANDES, E. (2017). A power analysis for knockoffs under Gaussian designs. *arXiv*.

WEINSTEIN, A., SU, W. J., BOGDAN, M., BARBER, R. F. and CANDES, E. J. (2020). A Power Analysis for Knockoffs with the Lasso. *arXiv*.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.
ZHONG, Y., KUFFNER, T. and LAHIRI, S. (2021). Conditional Randomization Rank Test. *arXiv* **1** 1–47.