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

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

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

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

Boldface (regular) font indicates population (sample) quantities. In a high-dimensional regression setting, H_{0n} is a model-agnostic way of formulating the null hypothesis that predictor X is unimportant in the regression of Y on (X, Z) (**?**). In a causal inference setting with treatment X, outcome Y, observed confounders Z, and no unobserved confounders, H_{0n} is the null hypothesis of no causal effect of X on Y (**?**).

---

[*]These authors contributed equally to this work.

As **?** 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 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* (**?**), which states that $\mathcal{L}_n(X|Z)$ is known exactly. Under the MX assumption, **?** 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 $\mathcal{L}_n(X|Z)$. Outside the context of randomized controlled experiments (**??**), the MX assumption is an approximation (**???**). In genome-wide association studies, a realistic parametric distribution can be postulated for this conditional law (**?**), 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 (**????????**). 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 $\mathcal{L}_n(X|Z)$ 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 (**?**), and the *generalized covariance measure (GCM) test*, a prototypical double regression approach to CI testing (**?**). Both tests involve learning $\mathcal{L}_n(X \mid Z)$ and $\mathcal{L}_n(Y \mid Z)$ in sample. Our main contributions are outlined next:

1. **The dCRT with $\mathcal{L}_n(X \mid Z)$ learned in sample can have poor Type-I error control if $\mathcal{L}_n(Y \mid Z)$ is learned poorly.** If $\mathcal{L}(X \mid Z)$ is known exactly, then the dCRT has finite-sample Type-I error control regardless of $\mathcal{L}(Y \mid Z)$ or the quality of its estimate. This is no longer the case once $\mathcal{L}(X \mid Z)$ 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 $\mathcal{L}_n(X \mid Z)$ can be compensated for by better approximations of $\mathcal{L}_n(Y \mid Z)$.** The MX assumption shifts the modeling burden entirely from $\mathcal{L}_n(Y \mid Z)$ to $\mathcal{L}_n(X \mid Z)$. When the latter is fit in sample, shifting the modeling burden partially back towards $\mathcal{L}_n(Y \mid Z)$ 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 X and 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 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). 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 (**??**) 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).

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 (**?**) and the CRT (**?**) 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, **?** 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 (**?**). 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, **?** 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 (**???**) probes the robustness of MX knockoffs with imperfect covariate distribution for a variety of specific test statistics and covariate distributions, with **??** 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 (**?**). 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, **?** proposed a variant of the CRT based on conditioning on an approximate sufficient statistic. Another method, the *conditional permutation test* (**?**), 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* (**?**) 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 (**????????**), although only **???** do not assume that $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is known exactly (the MX assumption). Beyond calculating power against certain alternatives, **?** and **?** 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

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

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, **?**), 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 $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$.

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

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

3 $\quad$ Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^n \mathcal{L}_n(X_i|Z_i)$ and compute $T_n(\widetilde{X}^{(m)}, Y, Z)$;

4 **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, **?**) 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 (**?**) and genomics (**?**) 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

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

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.

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

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

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

4 $\quad$ Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^{n} \widehat{\mathcal{L}}_n(X_i|Z_i)$ and compute

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

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 (**?**), defined as

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

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:

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

It controls Type-I error if the following in-sample mean-squared error quantities are small (**?**):

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

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

In particular, **?** require that

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

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

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

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 **?** for a discussion of this distinction.

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 (**?**) 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

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

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

$$\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z) \equiv \mathbb{1}(T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > \mathbb{Q}_{1-\alpha}[T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z])$$

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

$$\equiv \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}}T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > C_n^{\widehat{\text{dCRT}}}(X, Y, Z)\right).$$

The conditional $1 - \alpha$ quantile $C_n^{\widehat{\mathrm{dCRT}}}(X, Y, Z)$ 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 $\widehat{\mathrm{dCRT}}$. 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

$$(10) \qquad \mathbb{Q}_\alpha[W \mid \mathcal{F}] \equiv \inf\{t : \mathbb{P}[W \leq t \mid \mathcal{F}] \geq \alpha\}.$$

One would expect, based on the central limit theorem, that the conditional distribution of the ratio $T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z)/\widehat{S}_n^{\widehat{\mathrm{dCRT}}}$ 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

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

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 8), 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:*

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

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

If the conditional Lyapunov condition

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

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

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

and therefore

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

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

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

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.*

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

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

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

This result extends **?**, 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 $[\mathcal{L}_n(Y|Z)]$" (**?**). 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 $\mathcal{L}_n(X|Z)$ 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 $\mathcal{L}_n(Y|Z)$ (**?**). By contrast, in this section we show that when $\mathcal{L}_n(X|Z)$ 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

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

Suppose we fit $\mathcal{L}_n(X|Z)$ via a ridge regression while using the trivial estimate $\widehat{\mu}_{n,y}(Z) \equiv 0$ for $\mathbb{E}[Y|Z]$. 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 $\widehat{\mathcal{L}}_n(X|Z) = N(Z^T \widehat{\beta}_n, 1)$ and $\widehat{\mu}_{n,y}(Z) \equiv 0$. In this case, the normality of $\widehat{\mathcal{L}}_n(X|Z)$ 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:

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

On the other hand, it is easy to derive that

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

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

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

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.

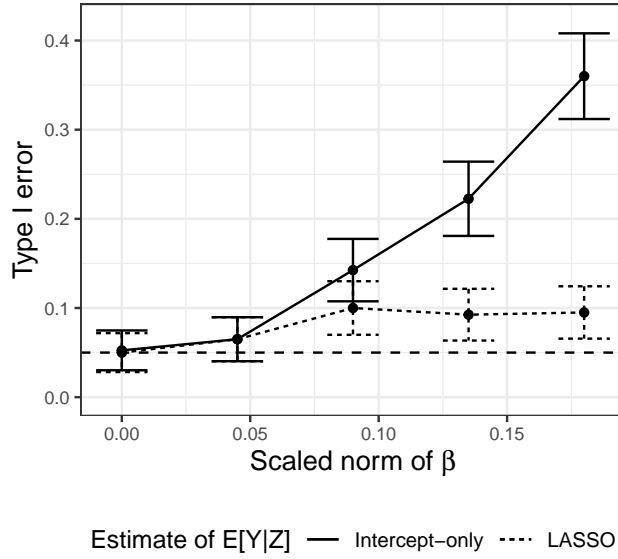Estimate of E[Y|Z] —— Intercept–only ···· LASSO

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.

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 **?** (see the discussion after Theorem 3). Similar phenomena have been noted in the contexts of causal inference (**?**) and doubly robust estimation (**??**); 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 (**??**). If the 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 (**?**). 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 $\text{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ is consistent in the following sense:

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

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

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

where

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

Furthermore, we denote by

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

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*

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

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

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

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

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

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. $\text{Var}_{\widehat{\mathcal{L}}_n}[X_i | Z_i] \equiv (X_i - \widehat{\mu}_{n,x}(Z_i))^2$;
2. $\text{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X} | \boldsymbol{Z}] \equiv f(\widehat{\mu}_{n,x}(\boldsymbol{Z}))$, *if*
   - $\text{Var}_{\mathcal{L}_n}[\boldsymbol{X} | \boldsymbol{Z}] = f(\mu_{n,x}(\boldsymbol{Z}))$ *for f Lipschitz on domain $\cup_{n=1}^{\infty} \text{Conv}(\text{supp}(\mathcal{L}_n(\boldsymbol{X})))$ and* $\text{supp}(\widehat{\mu}_{n,x}(\boldsymbol{Z})) \subseteq \text{Conv}(\text{supp}(\mathcal{L}_n(\boldsymbol{X})))$ *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 $\text{Conv}(\mathcal{L}_n(\boldsymbol{X}))$. 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 $\mathcal{L}'_n$:*

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

and therefore have the same asymptotic power:

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

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

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

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 $\widehat{\mathrm{dCRT}}$ 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 **?**. 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 $\widehat{\mathrm{dCRT}}$ test statistic. We explore this possibility in Appendix A.

Our double robustness result for the dCRT evokes the double robustness result proved for a conditional variant of MX knockoffs by **?**. 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 **?** finds that conditional knockoffs are *model doubly robust* (**?**). 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 $\widehat{\mathrm{dCRT}}$ 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 $Y|Z$..." (**?**). This statement is in reference to the worst-case performance of the CRT across all possible test statistics (**?**). 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 $\widehat{\mathrm{dCRT}}$ will control Type-I error (Corollary 3).

5. GCM test is optimal against certain alternatives. We have shown that, in large samples, the $\widehat{\mathrm{dCRT}}$ 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 (**???**) 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

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

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

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

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

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

We leave the dependence of $\theta_n(h)$ on g_0 implicit. Next, we define asymptotic optimality against such local alternatives following **?**:

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

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

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

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

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

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

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:*

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

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

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

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

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

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

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 **?** 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 **?**, 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

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

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 $k(x,y) = \min\{x,y\}$ (**?**, 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 **?**, Section 4. Using **?**, 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 $\mathcal{L}_{\theta_n(h)}(X, Y, Z)$ given by

$$\mathcal{L}_{\theta_n(h)}(Z) \equiv \mathrm{Unif}([0, 1]);$$

(43)
$$\mathcal{L}_{\theta_n(h)}(X|Z) \equiv N(\mu_{0x}(Z), 1);$$

$$\mathcal{L}_{\theta_n(h)}(Y|X, Z) \equiv N(Xh_\beta/\sqrt{n} + \mu_{0y}(Z) + h_g(Z)/\sqrt{n}, 1).$$

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 (**?**, Theorem 11); it is also optimal against local alternatives.

5.3. *Discussion of Theorem 3.* Theorem 3 states that the GCM test of **?** 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 **?** have already noted the connection between the GCM test and nonparametric estimation of the expected conditional covariance between X and Y given Z. Our result complements another line of work on minimax optimality for conditional independence testing (**???**). 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 (**?**) and the other in the context of model-X knockoffs (**?**).

Theorem 3 closely parallels results on estimation in semiparametric regression (**?????????**). It follows from **??** 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 (**???**).

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).

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 **?**, 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 Z. Indeed, semiparametric optimality theory is predominantly low-dimensional. A notable exception is the work of **?**, 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 (**??**), model-X knockoffs (**?**), and the GCM test (**?**) 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 **?**.

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: **????** found broad robustness to misspecification of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ while **?** 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-v1.

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 (**?**). 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 (**????**), the latter studying the dCRT specifically. On the other hand, the numerical simulations of **?** 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:

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

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:

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

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 **?** 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 **???** (Appendix F.1). The former two papers are framed in the variable selection context, where several explanatory variables $\boldsymbol{W}_j$ are considered, and the hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{-j}$ is tested for each j. Therefore, $\boldsymbol{X} \equiv \boldsymbol{W}_j$ for each j. On the other hand, **?** considered a conditional independence testing framework, where $\boldsymbol{X}$ was a single variable of interest.

For the data-generating models used by **??**, we evaluate the Type-I error of the marginal GCM test for each hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{-j}$, 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 **?** (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 **?** and 0.5 by **?**) 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 **??**.

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:

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

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
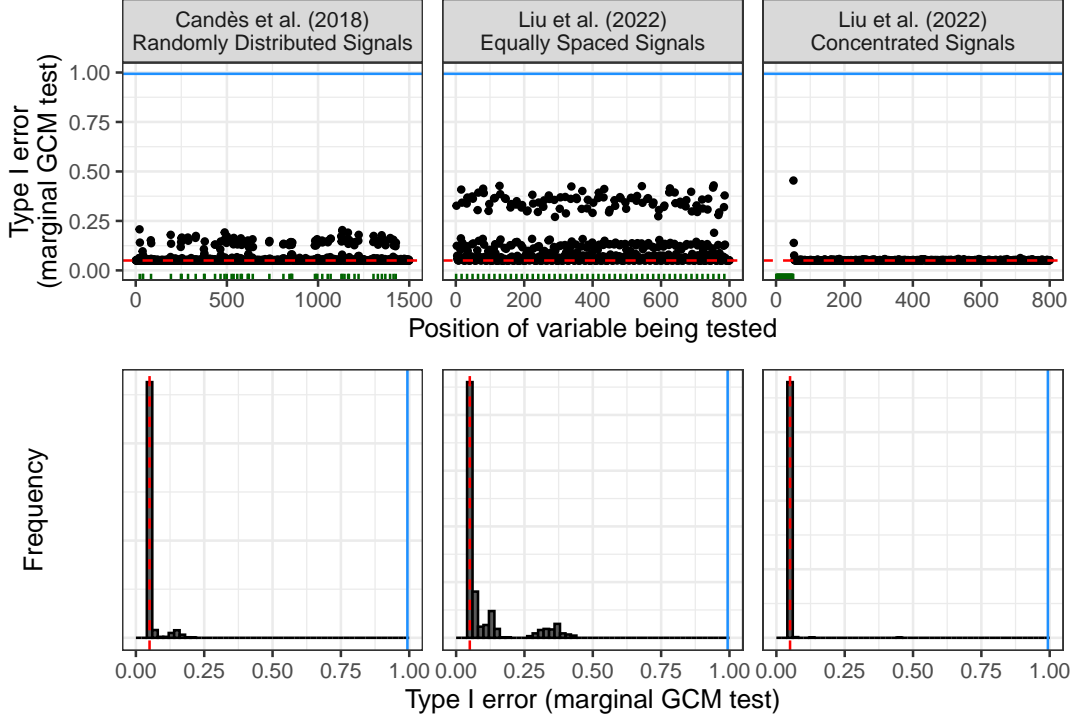
Fig 2: Comparing the marginal associations between X and Y in the robustness simulations of **???** (Appendix F.1). Top: Type-I error of the marginal GCM test as a function of the position of null variables with respect to the non-null variables (represented as green ticks). Bottom: Histograms of the 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 **?**, and the dashed red line the nominal Type-I error level of the marginal GCM test (0.05).

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.

n	p	s	ρ	θ (null)	ν (null)	θ (alt)	ν (alt)
100	100	**5**	0	0	0	0	$\nu_{\max}/2$
200	200	10	0.2	0	$\nu_{\max}/4$	$\theta_{\max}/4$	$\nu_{\max}/2$
400	**400**	20	**0.4**	0	$\nu_{\max}/2$	$\theta_{\max}/2$	$\nu_{\max}/2$
800	800	40	0.6	0	$3\nu_{\max}/4$	$3\theta_{\max}/4$	$\nu_{\max}/2$
1600	1600	80	0.8	0	$\nu_{\max}$	$\theta_{\max}$	$\nu_{\max}/2$

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 (**?**), 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 X and Z but not of Y. The methodology (specifically, "Maxway$_{\text{in}}$ example 1") proceeds—roughly—by fitting $\widehat{\mathcal{L}}_n(X|Z)$ on the unlabeled data via the post-lasso (i.e. selecting active variables via the lasso and then refitting via ordinary least squares, **?**), fitting $\widehat{\mu}_{ny}(Z)$ 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 **?**, 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 **?**. However, in the semi-supervised setting we use all of the available data on (X, Z) (i.e. both unlabeled and labeled data) to fit $\mathcal{L}_n(X|Z)$. By contrast, **?** used only the unlabeled data to learn $\mathcal{L}_n(X|Z)$ 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 **?** is attributable to poor estimation of $\mu_n(X|Z)$ and/or $\mu_n(Y|Z)$ 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 **?**, 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

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

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

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.

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 lower and upper critical values is necessary, as the null distribution may not be symmetric and centered at zero (**?**).

6.3. *Simulation results.* We conducted simulations for Gaussian and binary models for the response 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. 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.

Next we list the main conclusions regarding Type-I error based on the results in Figures 3 (Gaussian supervised), 8 (Gaussian semi-supervised), 10 (binary supervised), and 12 (binary semi-supervised):

- 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 10). 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.
- 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 Z out of the response Y; see also Appendix F.2.
- 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 **?**), 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 (X, Z) to fit $\mathcal{L}_n(X|Z)$, and therefore does not adjust for Z as accurately.
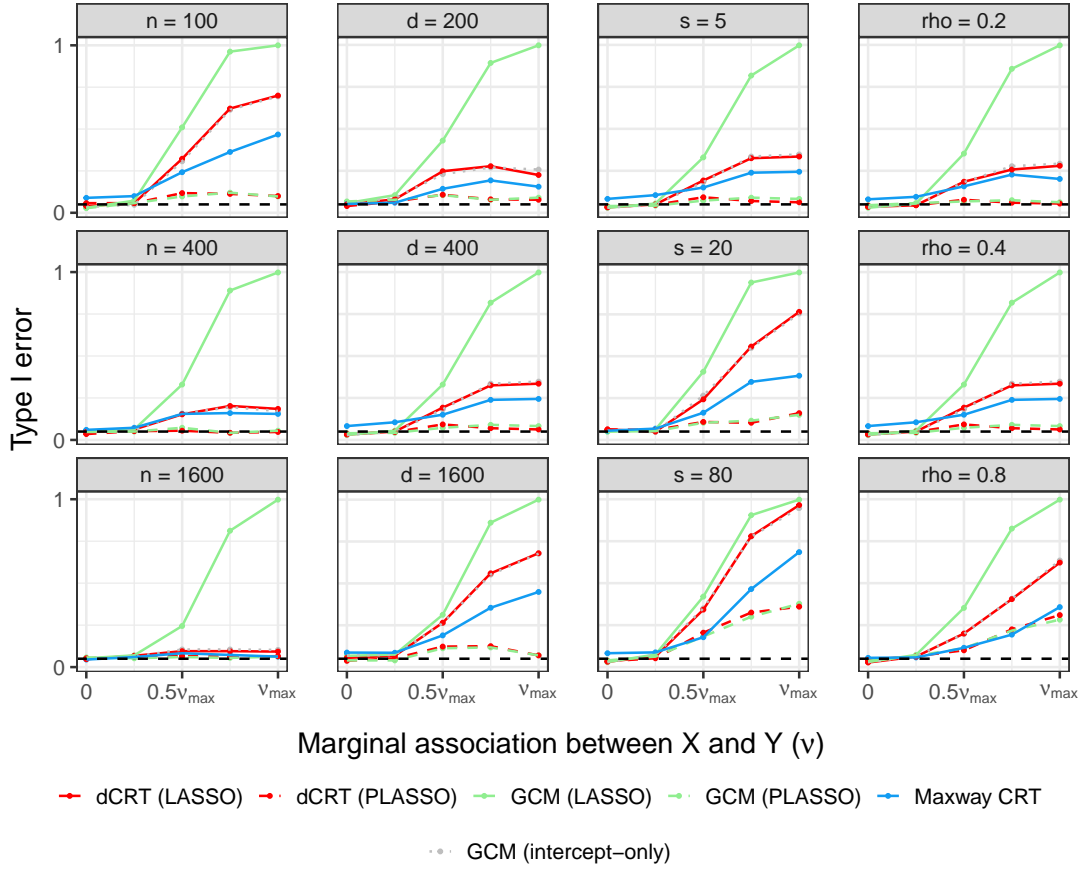
Fig 3: Type I error control for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the marginal association strength ν in each subplot. Each point is the average of 400 Monte Carlo replicates. All the standard errors are less than 0.026.

Next, we list the main conclusions regarding power based on the results in Figures 4 (Gaussian supervised), 9 (Gaussian semi-supervised), 11 (binary supervised), and 13 (binary semi-supervised):

- 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.
- Across Gaussian and binary supervised simulation settings (Figures 7 and 11), 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 9 and 13), Maxway CRT has power comparable to or better than those of the post-lasso-based meth-

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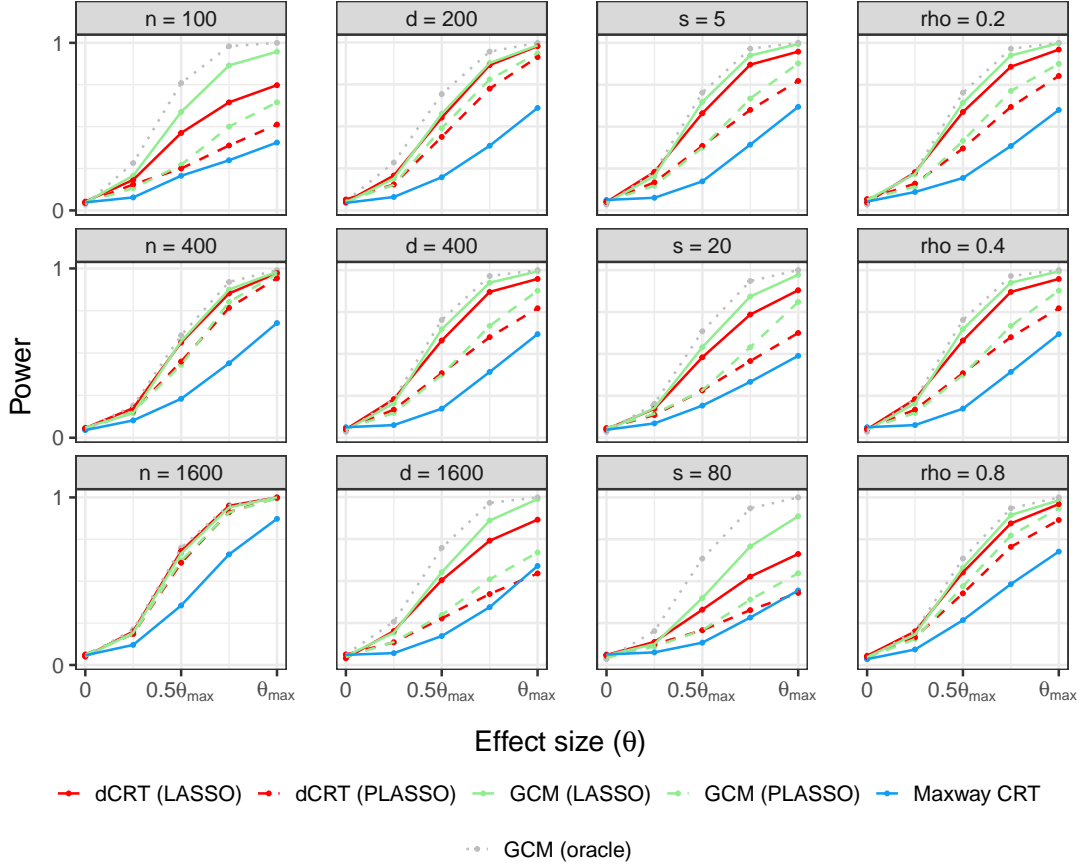

Fig 4: Power for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the signal strength θ in each subplot. Each point is the average of 400 Monte Carlo replicates. All the standard errors are less than 0.026.

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.

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 (?) 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 (**?**). A doubly robust version of the dCRT has also been recently proposed (the Maxway CRT; **?**), 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 (**?**), 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 10 and 14), 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 **?**, 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 6), 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 **?**, 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 X and 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 $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$ 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 $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$ depend on at most $s = o(\sqrt{n}/\log(p))$ of the coordinates of Z. Such assumptions are common in other lines of work on high-dimensional / semiparametric / doubly robust inference, including the debiased lasso (**?????**) and doubly robust causal inference (**??**). On the other hand, consistent estimates are typically not available in the regime when n, p, and s grow proportionally (**?**), causing a failure in traditional debiased estimates (**?**). 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 (**?**). 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.

APPENDIX A: THE $\widehat{\text{dCRT}}$ WITH GCM NORMALIZATION

As an alternative to the $\widehat{\text{dCRT}}$, we consider the $\widehat{\text{ndCRT}}$. This procedure is based on a normalized statistic that coincides exactly with the GCM statistic:

$$T_n^{\widehat{\text{ndCRT}}}(X,Y,Z) \equiv \frac{1}{\widehat{S}_n^{\text{GCM}}} T_n^{\widehat{\text{dCRT}}}(X,Y,Z) \equiv T_n^{\text{GCM}}(X,Y,Z).$$

The only difference with GCM is that the critical value is given by conditional resampling rather than a normal quantile:

(47)
$$\phi_n^{\widehat{\text{ndCRT}}}(X,Y,Z) \equiv \mathbb{1}(T_n^{\widehat{\text{ndCRT}}}(X,Y,Z) > \mathbb{Q}_{1-\alpha}[T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z])$$
$$\equiv \mathbb{1}\left(T_n^{\widehat{\text{ndCRT}}}(X,Y,Z) > C_n^{\widehat{\text{ndCRT}}}(X,Y,Z)\right).$$

Here,

$$T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \equiv \frac{T_n^{\widehat{\text{dCRT}}}(\widetilde{X},X,Y,Z)}{S_n^{\text{GCM}}(\widetilde{X},X,Y,Z)},$$

where

$$(S_n^{\text{GCM}}(\widetilde{X},X,Y,Z))^2 \equiv \widehat{\text{Var}}\{(\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))\}$$

and $T_n^{\widehat{\text{dCRT}}}(\widetilde{X},X,Y,Z)$ is as defined in equation (5).

THEOREM 4. *Let $\mathcal{L}_n$ be a sequence of laws such that the nondegeneracy conditions* (NDG1) *and* (NDG2) *and the conditional Lyapunov condition* (Lyap-1) *hold. Then,*

(48)
$$T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z \xrightarrow{d,p} N(0,1)$$

and therefore

(49)
$$C_n^{\widehat{\text{ndCRT}}}(X,Y,Z) \equiv \mathbb{Q}_{1-\alpha}\left[T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z\right] \to z_{1-\alpha}$$

and the $\widehat{\text{ndCRT}}$ and GCM tests are equivalent:

(50)
$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{ndCRT}}}(X,Y,Z) = \phi_n^{\text{GCM}}(X,Y,Z)] = 1.$$

COROLLARY 5. *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 the nondegeneracy conditions* (NDG1) *and* (NDG2), *the conditional Lyapunov condition* (Lyap-1), *and the assumptions* (SP1) *and* (SP2). *Then, the $\widehat{\text{ndCRT}}$ has asymptotic Type-I error control over $\mathscr{L}_n^0 \cap \mathscr{R}_n$ in the sense of the definition* (28).

Comparing Corollary 5 to Corollary 3, we see that the $\widehat{\text{ndCRT}}$ controls Type-I error under weaker assumptions than required for the $\widehat{\text{dCRT}}$; in particular the variance of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ need not be estimated with any degree of accuracy.

APPENDIX B: CONDITIONAL CONVERGENCE RESULTS

The proofs of our theoretical results rely on the conditional counterparts of several standard convergence theorems. In this section, we state these conditional convergence theorems. We defer their proofs to Appendix G.

First we define a notion of conditional convergence in probability, analogous to our definition of conditional convergence in distribution (Definition 1).

DEFINITION 3.   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 probability to a constant c conditionally on $\mathcal{F}_n$ if W_n converges in distribution to the delta mass at c conditionally on $\mathcal{F}_n$ (recall Definition 1). We denote this convergence by $W_n \mid \mathcal{F}_n \xrightarrow{p,p} c$. In symbols,

$$(51) \qquad W_n \mid \mathcal{F}_n \xrightarrow{p,p} c \quad \text{if} \quad W_n \mid \mathcal{F}_n \xrightarrow{d,p} \delta_c.$$

Now we are ready to state the conditional convergence results.

B.1. Statements.   For the sake of all results below, let $\mathcal{F}_n$ be a sequence of σ-algebras.

THEOREM 5 (Conditional Polya's theorem).   *Let W_n be a sequence of random variables. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W with continuous CDF, then*

$$(52) \qquad \sup_{t \in \mathbb{R}} |\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| \xrightarrow{p} 0.$$

THEOREM 6 (Conditional Slutsky's theorem).   *Let W_n be a sequence of random variables. Suppose a_n and b_n are sequences of random variables such that $a_n \xrightarrow{p} 1$ and $b_n \xrightarrow{p} 0$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W with continuous CDF, then*

$$(53) \qquad a_n W_n + b_n \mid \mathcal{F}_n \xrightarrow{d,p} W.$$

THEOREM 7 (Conditional law of large numbers).   *Let W_{in} be a triangular array of random variables, such that W_{in} are independent conditionally on $\mathcal{F}_n$ for each n. If for some $\delta > 0$ we have*

$$(54) \qquad \frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] \xrightarrow{p} 0,$$

then

$$(55) \qquad \frac{1}{n} \sum_{i=1}^{n} (W_{in} - \mathbb{E}[W_{in} \mid \mathcal{F}_n]) \mid \mathcal{F}_n \xrightarrow{p,p} 0.$$

The condition (54) is satisfied when

$$(56) \qquad \sup_{1 \leq i \leq n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] = o_p(n^{\delta}).$$

As a corollary of Theorem 7, if we choose $\mathcal{F}_n = \{\varnothing, \Omega\}$, we are able to obtain the following version of the weak law of large numbers for triangular arrays.

COROLLARY 6 (Unconditional weak law of large numbers).   *Let W_{in} be a triangular array of random variables, such that W_{in} are independent for each n. If for some $\delta > 0$ we have*

$$(57) \qquad \frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta}] \to 0,$$

then

$$(58) \qquad \frac{1}{n} \sum_{i=1}^{n} (W_{in} - \mathbb{E}[W_{in}]) \xrightarrow{p} 0.$$

The condition (57) *is satisfied when*

$$(59) \qquad \sup_{1 \le i \le n} \mathbb{E}[|W_{in}|^{1+\delta}] = o(n^\delta).$$

THEOREM 8 (Conditional central limit theorem).   *Let W_{in} be a triangular array of random variables, such that for each n, W_{in} are independent conditionally on $\mathcal{F}_n$. Define*

$$(60) \qquad S_n^2 \equiv \sum_{i=1}^{n} \mathrm{Var}[W_{in} \mid \mathcal{F}_n],$$

and assume $0 < \mathrm{Var}[W_{in}|\mathcal{F}_n] < \infty$ almost surely for all $i = 1, \dots, n$ and for all $n \in \mathbb{N}$. If for some $\delta > 0$ we have

$$(61) \qquad \frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^{2+\delta} \mid \mathcal{F}_n] \xrightarrow{p} 0,$$

then

$$(62) \qquad \frac{1}{S_n} \sum_{i=1}^{n} (W_{in} - \mathbb{E}[W_{in} \mid \mathcal{F}_n]) \mid \mathcal{F}_n \xrightarrow{d,p} N(0,1).$$

LEMMA 1 (Conditional convergence implies quantile convergence).   *Let W_n be a sequence of random variables and $\alpha \in (0,1)$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W whose CDF is continuous and strictly increasing at $\mathbb{Q}_\alpha[W]$, then*

$$(63) \qquad \mathbb{Q}_\alpha[W_n \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{Q}_\alpha[W].$$

B.2. Discussion.   The above definitions and results on conditional convergence are not particularly surprising, and related results are present in the existing literature. Nevertheless, we have not found any of the above results stated in the literature in exactly this form. Here we discuss the relationships of our definitions and results with existing ones.

Notions of conditional convergence in probability and in distribution have been explicitly defined by **?**. However, these notions require a single conditioning σ-algebra as well as almost sure convergences of conditional probabilities, whereas in Definitions 1 and 3 we allow the conditioning σ-algebra to change with n and for the conditional probabilities to converge in probability. Our Definition 1 can be viewed as formalizing the notion of conditional convergence in distribution implicitly used by **?**. Related notions of conditional convergence in distribution allowing for changing conditioning σ-algebra are present implicitly in the works of **?** and **?**, though these are based on the convergence of conditional characteristic functions as opposed to conditional cumulative distribution functions.

Turning to the convergence results themselves, we were not able to find conditional Polya's theorem (Theorem 5) in the literature. Conditional Slutsky's theorem (Theorem 6) is a generalization of **?**, Lemma 5 to the case when a_n is not necessarily independent of $\mathcal{F}_n$ and $b_n \ne 0$. Versions of the conditional law of large numbers are given by **?** and **?**, but these involve a single conditioning σ-algebra and do not allow for triangular arrays, unlike Theorem 7. Remarkably, we could not find even the unconditional triangular array law of large numbers (Corollary 6) in the literature; existing results either assume a second-moment condition or use truncation (**?**, Theorems 2.2.4 and 2.2.6, respectively) instead of a $1 + \delta$ moment condition or are not applicable to triangular arrays (**?**, Lemma 19). As for central limit theorems, **???** give non-triangular array versions of the conditional central limit theorem that require a single conditioning σ-algebra, unlike Theorem 8. Versions of the conditional central limit theorem appropriate for varying conditioning σ-algebras and triangular arrays are

given by **??**, those these involve different notions of conditional convergence in distribution. Results similar to Theorem 8 for $\delta = 1$ are presented in a recent line of work on sample-splitting-based inference (**???**); these can be proved via the Berry-Esseen theorem. Finally, we note that our result that conditional convergence in distribution implies in-probability quantile convergence (Lemma 1) is a generalization of **?**, Lemma 3 to general conditioning σ-algebras.

<div align="center">APPENDIX C: PROOFS FOR SECTION 2</div>

C.1. Auxiliary lemmas.

LEMMA 2. *Let W_n be a sequence of nonnegative random variables and let $\mathcal{F}_n$ be a sequence of σ-algebras. If $\mathbb{E}[W_n \mid \mathcal{F}_n] \xrightarrow{p} 0$, then $W_n \xrightarrow{p} 0$.*

PROOF. For any $\epsilon > 0$, we have

$$\mathbb{P}[W_n \geq \epsilon] = \mathbb{P}[W_n \wedge \epsilon \geq \epsilon] \tag{64}$$

$$\leq \epsilon^{-1} \mathbb{E}[W_n \wedge \epsilon] \tag{65}$$

$$= \epsilon^{-1} \mathbb{E}[\mathbb{E}[W_n \wedge \epsilon \mid \mathcal{F}_n]] \tag{66}$$

$$\leq \epsilon^{-1} \mathbb{E}[\mathbb{E}[W_n \mid \mathcal{F}_n] \wedge \epsilon] \to 0, \tag{67}$$

where the last convergence is due to bounded convergence theorem and the assumption $\mathbb{E}[W_n \mid \mathcal{F}_n] \xrightarrow{p} 0$. □

LEMMA 3 (Asymptotic equivalence of tests). *Consider two hypothesis tests based on the same test statistic $T_n(X, Y, Z)$ but different critical values:*

$$\phi_n^1(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > C_n(X, Y, Z)); \quad \phi_n^2(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > z_{1-\alpha}).$$

If the critical value of the first converges in probability to that of the second:

$$C_n(X, Y, Z) \xrightarrow{p} z_{1-\alpha} \tag{68}$$

and the test statistic does not accumulate near the limiting critical value:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - z_{1-\alpha}| \leq \delta] = 0, \tag{69}$$

then the two tests are asymptotically equivalent:

$$\lim_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^1(X, Y, Z) = \phi_n^2(X, Y, Z)] = 1. \tag{70}$$

PROOF. Note that for any $\delta > 0$, we have

$$\mathbb{P}_{\mathcal{L}_n}\left[\phi_n^1(X, Y, Z) \neq \phi_n^2(X, Y, Z)\right]$$

$$= \mathbb{P}_{\mathcal{L}_n}\left[\min(z_{1-\alpha}, C_n) < T_n \leq \max(z_{1-\alpha}, C_n)\right]$$

$$= \mathbb{P}_{\mathcal{L}_n}\left[\min(z_{1-\alpha}, C_n) < T_n \leq \max(z_{1-\alpha}, C_n), |C_n - z_{1-\alpha}| \leq \delta\right] \tag{71}$$

$$\quad + \mathbb{P}_{\mathcal{L}_n}\left[\min(z_{1-\alpha}, C_n) < T_n \leq \max(z_{1-\alpha}, C_n), |C_n - z_{1-\alpha}| > \delta\right]$$

$$\leq \mathbb{P}_{\mathcal{L}_n}\left[|T_n - z_{1-\alpha}| \leq \delta\right] + \mathbb{P}_{\mathcal{L}_n}\left[|C_n - z_{1-\alpha}| > \delta\right].$$

To justify the last step, suppose without loss of generality that $z_{1-\alpha} \leq C_n$. Then note that if $z_{1-\alpha} < T_n \leq C_n$ and $C_n - z_{1-\alpha} \leq \delta$ then

$$|T_n - z_{1-\alpha}| = T_n - z_{1-\alpha} \leq C_n - z_{1-\alpha} \leq \delta.$$

Taking a limsup on both sides in equation (71) and using the assumed convergence (68), we find that

$$\limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n} \left[\phi_n^1(X,Y,Z) \neq \phi_n^2(X,Y,Z) \right]$$

$$\leq \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n} \left[|T_n(X,Y,Z) - z_{1-\alpha}| \leq \delta \right] + \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n} \left[|C_n(X,Y,Z) - z_{1-\alpha}| > \delta \right]$$

$$= \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n} \left[|T_n(X,Y,Z) - z_{1-\alpha}| \leq \delta \right].$$

Letting $\delta \to 0$ and using our assumption (69), we arrive at the claimed asymptotic equivalence. This completes the proof. $\qquad\square$

C.2. Proofs of main results.

PROOFS OF THEOREM 1 AND COROLLARY 1. We prove instead the stronger Theorem 9 and Corollary 7 below. $\qquad\square$

THEOREM 9. *Let $\mathcal{L}_n$ be a sequence of laws and $\widehat{\mathcal{L}}_n$ be a sequence of estimates. Suppose either of the following two sets of assumptions is satisfied:*

1. *The nondegeneracy conditions (NDG1) and (NDG2) and the conditional Lyapunov condition (Lyap-1) hold.*
2. *The assumptions of Theorem 2 hold.*

Then, the normalized $T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z)$ converges conditionally to a normal distribution (12) and therefore the $\widehat{\mathrm{dCRT}}$ critical value $C_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)$ converges to $z_{1-\alpha}$ (13).

PROOF. It suffices to prove the conditional convergence in distribution (12), as the convergence of the critical value (13) follows from Lemma 1 because the normal distribution has continuous and strictly increasing CDF. We prove the conditional convergence (12) for each of the two sets of assumptions.

Assumption 1. We proceed by applying the conditional CLT (Theorem 8) with

$$(72) \qquad W_{in} \equiv \frac{1}{\sqrt{n}}(\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))$$

and $\mathcal{F}_n \equiv \sigma(X,Y,Z)$. To verify the assumptions of the conditional CLT, note first that W_{in} are independent conditionally on $\mathcal{F}_n$ by construction and satisfy $0 < \mathrm{Var}[W_{in} \mid \mathcal{F}_n] < \infty$ by the nondegeneracy assumption (NDG2). Next, recalling definition (9), we have

$$S_n^2 \equiv \sum_{i=1}^{n} \mathrm{Var}_{\widehat{\mathcal{L}}_n}[W_{in} \mid X,Y,Z] \equiv (\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2,$$

so that

$$\frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}_{\widehat{\mathcal{L}}_n}[|W_{in}|^{2+\delta} \mid X,Y,Z]$$

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

This quantity converges to zero in probability due to the nondegeneracy condition (NDG1) and the Lyapunov condition (Lyap-1). Hence, the conditional CLT gives the desired conditional convergence (12).

Assumption 2. We begin by decomposing $T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z)$:

$$(73) \qquad T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i))$$

$$(74) \qquad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))$$

$$(75) \qquad \equiv I_n - J_n.$$

We claim that $J_n \xrightarrow{p} 0$. Indeed, from $\mathbb{E}[J_n \mid X, Y, Z] = 0$ and the assumption $\widehat{E}_{n,y}' \xrightarrow{p} 0$ (SP1')
it follows

$$\mathbb{E}[J_n^2 \mid X, Y, Z] = \text{Var}[J_n \mid X, Y, Z] \equiv (\widehat{E}_{n,y}')^2 \xrightarrow{p} 0.$$

The following lemma which which appears as Lemma 2 is particularly useful now:

LEMMA 4. *Let W_n be a sequence of nonnegative random variables and let $\mathcal{F}_n$ be a sequence of σ-algebras. If $\mathbb{E}[W_n \mid \mathcal{F}_n] \xrightarrow{p} 0$, then $W_n \xrightarrow{p} 0$.*

Hence by Lemma 2 we have $J_n^2 \xrightarrow{p} 0$, so that $J_n \xrightarrow{p} 0$, as claimed. Next, we claim that an appropriately rescaled I_n converges conditionally to $N(0, 1)$. To this end, we apply the conditional CLT (Theorem 8) with

$$(76) \qquad W_{in} \equiv \frac{1}{\sqrt{n}} (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i))$$

and $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. To verify the assumptions of the conditional CLT, note first that W_{in} are independent conditionally on $\mathcal{F}_n$ by construction and satisfy $0 < \text{Var}[W_{in} \mid \mathcal{F}_n] < \infty$ by the nondegeneracy assumption (NDG2). Next, observe that

$$S_n^2 \equiv \sum_{i=1}^{n} \text{Var}[W_{in} \mid X, Y, Z] = \frac{1}{n} \sum_{i=1}^{n} \text{Var}_{\widehat{\mathcal{L}}_n}[X_i | Z_i](Y_i - \mu_{n,y}(Z_i))^2 \equiv (S_n^{\widehat{\text{dCRT}}})^2,$$

so that

$$\frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid X, Y, Z]$$

$$= \frac{1}{(S_n^{\widehat{\text{dCRT}}})^{2+\delta}} \cdot \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} |Y_i - \mu_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z].$$

Since the first factor is stochastically bounded (conclusion (95) from Lemma 8), it suffices to show that the second factor converges to zero in probability. To this end, by Lemma 2 it suffices to note that $\mathcal{L}_n \in \mathscr{L}_n^0$ and the Lyapunov assumption (Lyap-2) give

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

$$(77) \qquad = \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{L}_n}[|Y_i - \mu_{n,y}(Z_i)|^{2+\delta} | Z_i] \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z]$$

$$\xrightarrow{p} 0.$$

Therefore, we may apply the conditional CLT to obtain that

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

Furthermore, equation (94) from Lemma 8 gives $S_n^{\widehat{\text{dCRT}}}/\widehat{S}_n^{\widehat{\text{dCRT}}} \xrightarrow{p} 1$, so by conditional Slutsky's theorem (Theorem 6) we conclude that

$$(78) \qquad \frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) = \frac{S_n^{\widehat{\text{dCRT}}}}{\widehat{S}_n^{\widehat{\text{dCRT}}}} \left(\frac{I_n}{S_n^{\widehat{\text{dCRT}}}} - \frac{J_n}{S_n^{\widehat{\text{dCRT}}}} \right) \Bigg| X, Y, Z \xrightarrow{d,p} N(0,1),$$

as desired. $\qquad\square$

COROLLARY 7. *Under the assumptions of Theorem 9, if the non-accumulation condition (15) is satisfied then the $\widehat{\text{dCRT}}$ is asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test.*

PROOF. The desired equivalence is given by Lemma 3 with $T_n(X, Y, Z) \equiv (\widehat{S}_n^{\widehat{\text{dCRT}}})^{-1} T_n^{\widehat{\text{dCRT}}}$ and $C_n(X, Y, Z) \equiv C_n^{\widehat{\text{dCRT}}}(X, Y, Z)$, since ϕ_n^1 and ϕ_n^2 in the lemma statement reduce to $\widehat{\text{dCRT}}$ and the $\widehat{\text{MX(2)}}$ F-test, respectively. This lemma is applicable because the convergence of the critical value (68) is given by Theorem 9 and the non-accumulation condition (69) is assumed. $\qquad\square$

APPENDIX D: PROOFS FOR SECTION 4 AND APPENDIX A

For the sake of this section, we define

$$(79) \qquad s_n^2 \equiv \mathbb{E}_{\mathcal{L}_n}[\text{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]\text{Var}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]].$$

D.1. Auxiliary lemmas.

LEMMA 5 (Conditional Jensen inequality, **?**, Theorem 10.18). *Let W be a random variable and let ϕ be a convex function, such that W and $\phi(W)$ are integrable. For any σ-algebra $\mathcal{F}$, we have the inequality*

$$\phi(\mathbb{E}[W \mid \mathcal{F}]) \leq \mathbb{E}[\phi(W) \mid \mathcal{F}] \quad \text{almost surely.}$$

LEMMA 6. *Let W_n be a sequence of random variables and $\mathcal{F}_n$ a sequence of σ-algebras. If $W_n \mid \mathcal{F}_n \xrightarrow{p,p} 0$, then $W_n \xrightarrow{p} 0$.*

PROOF. Let $\epsilon > 0$. Because of the assumed conditional convergence in probability, we have

$$\mathbb{P}[|W_n| > \epsilon \mid \mathcal{F}_n] \xrightarrow{p} 0.$$

By the bounded convergence theorem, it follows that

$$\mathbb{P}[|W_n| > \epsilon] = \mathbb{E}[\mathbb{P}[|W_n| > \epsilon \mid \mathcal{F}_n]] \to 0,$$

from which the conclusion follows. $\qquad\square$

LEMMA 7.   *Consider a sequence of laws $\mathcal{L}_n \in \mathscr{L}_n^0$. Given assumption (SP2), we have*

$$(80) \qquad \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\operatorname{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0;$$

$$(81) \qquad \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2\operatorname{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0;$$

$$(82) \qquad \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_{\mathcal{L}_n}[X_i \mid Z_i](Y_i - \mu_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0.$$

Given additionally assumption (SP1), we also have

$$(83) \qquad \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{n,x}(Z_i))^2\operatorname{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0;$$

$$(84) \qquad \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_{\mathcal{L}_n}[X_i \mid Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0.$$

Given additionally the conditional Lyapunov condition (Lyap-2) and the variance consistency condition (21), we also have

$$(85) \qquad (S_n^{\widehat{\mathrm{dCRT}}})^2 - s_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \mu_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0.$$

Given additionally the assumption $\widehat{E}'_{n,y} \xrightarrow{p} 0$, we also have

$$(86) \qquad (\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2 - s_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0.$$

PROOF. We prove the convergence statements in order.

Proofs of statements (80), (81), (82).   These statements are consequences of the weak law of large numbers (Corollary 6). To verify the statement (80), we note that

$$\mathbb{E}_{\mathcal{L}_n}[\operatorname{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\operatorname{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]] = \mathbb{E}_{\mathcal{L}_n}[\operatorname{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\operatorname{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \equiv s_n^2,$$

and

$$\sup_{i,n} \mathbb{E}_{\mathcal{L}_n}[|\operatorname{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\operatorname{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]|^{1+\delta/2}]$$

$$= \sup_{n} \mathbb{E}_{\mathcal{L}_n}[|\operatorname{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\operatorname{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]|^{1+\delta/2}]$$

$$(87) \qquad \leq \sup_{n} \mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z})|^{2+\delta} \mid \boldsymbol{Z}]\mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta} \mid \boldsymbol{Z}]]$$

$$= \sup_{n} \mathbb{E}_{\mathcal{L}_n}[|(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))|^{2+\delta}]$$

$$< c_2 < \infty.$$

The inequality in the third line follows from the conditional Jensen inequality (Lemma 5), the equality in the fourth line follows from the conditional independence assumption, and the inequality in the fifth line follows from the $2 + \delta$ moment assumption (SP2). Hence we have verified the sufficient condition (59) for the WLLN, so the convergence (80) follows. Statements (81) and (82) can be verified with similar arguments.

Proofs of statements (83) *and* (84). We prove only the first, as the second will follow by symmetry. Given the convergence (81), the statement (83) will follow if we show that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

(88)
$$- \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$\equiv (E'_{n,x})^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$\xrightarrow{p} 0.$$

The convergence $E'_{n,x} \xrightarrow{p} 0$ is assumed (SP1). To verify the convergence of the second term, note first that conditional Hölder (Lemma 13) and the derivation (87) give

(89)
$$\sup_{n} s_n^2 \leq \sup_{n} \{\mathbb{E}_{\mathcal{L}_n}[|\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]|^{1+\delta/2}]\}^{2/(2+\delta)} < \infty,$$

which combined with the convergence (81) implies that

(90)
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] = O_p(1).$$

Therefore, by the Cauchy-Schwartz inequality we find that

$$\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)^2$$

$$\leq \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)$$

$$\equiv \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right) \cdot (E'_{n,x})^2$$

$$= O_p(1) \cdot o_p(1) = o_p(1).$$

This proves the convergence (88), which in turn implies the claimed convergence (83).

Proof of statement (85). Note that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-s_n^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]-\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]-s_n^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]+o_p(1),$$

where we used the variance consistency assumption (21) and the convergence result (80) to obtain the last line. Hence, it suffices to show that

$$(91)\qquad \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}[Y_i\mid Z_i]\right)\overset{p}{\to}0.$$

To this end, we apply the conditional WLLN (Theorem 7) with $\mathcal{F}_n=\sigma(X,Z)$ and

$$W_{in}=\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2.$$

We check the required $1+\delta$ moment condition (54):

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}[|W_{in}|^{1+\delta/2}\mid\mathcal{F}_n]$$

$$\equiv\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}\left[\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]^{1+\delta/2}\cdot|Y_i-\mu_{n,y}(Z_i)|^{2+\delta}\mid X,Z\right]$$

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

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

$$\overset{p}{\to}0.$$

The inequality in the third line follows from the conditional Jensen inequality (Lemma 5), the equality in the fourth line from the assumed conditional independence, and the convergence in the fifth line from the conditional Lyaponov assumption (Lyap-2). Therefore, the conditional WLLN gives

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i\mid Z_i]\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2|X,Z]\right)$$

$$\overset{p,p}{\longrightarrow}0,$$

where the equality follows from the assumed conditional independence. Since conditional convergence in probability implies unconditional convergence in probability (Lemma 6), this verifies the claimed convergence statement (91) and completes the proof of the statement (85).

Proof of statement (86). Given the convergence (85), it suffices to show that

$$(92) \quad \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \mu_{n,y}(Z_i))^2 \xrightarrow{p} 0$$

Given the assumption that $\widehat{E}'_{n,y} \xrightarrow{p} 0$, this statement's proof is analogous to that of statement (88), so we omit it for the sake of brevity. This completes the proof of the lemma. $\square$

LEMMA 8. *Define*

$$(93) \qquad (S_n^{\widehat{\mathrm{dCRT}}})^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i](Y_i - \mu_{n,y}(Z_i))^2.$$

Under the assumptions of Theorem 2, we have

$$(94) \qquad \frac{(S_n^{\widehat{\mathrm{dCRT}}})^2}{(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2} \xrightarrow{p} 1$$

and

$$(95) \qquad \mathbb{P}[(S_n^{\widehat{\mathrm{dCRT}}})^2 > \epsilon] \to 1 \quad \textit{for some } \epsilon > 0.$$

PROOF. The equivalence of variances (94) follows from the convergences (85) and (86) (Lemma 7), as well as the observation that conditional independence and the assumption (SP2) imply

$$(96) \qquad \inf_n s_n^2 = \inf_n \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))^2(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))^2] > 0.$$

The stochastic boundedness from below (95) follows from the latter fact and the convergence (85). $\square$

LEMMA 9. *If the nondegeneracy conditions* (NDG1) *and* (NDG2) *and the conditional Lyapunov condition* (Lyap-1) *hold, then the variance estimates* $(\widehat{S}_n^{\mathrm{GCM}})^2$ *and* $(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2$ *are equivalent under resampling:*

$$(97) \qquad \frac{(\widehat{S}_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z))^2}{(\widehat{S}_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z))^2} \xrightarrow{p} 1.$$

PROOF. Define

$$W_{in} \equiv (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)),$$

so that

$$(\widehat{S}_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z))^2 \equiv \frac{1}{n}\sum_{i=1}^{n}W_{in}^2 - \left(\frac{1}{n}\sum_{i=1}^{n}W_{in}\right)^2.$$

First we claim that $\frac{1}{n}\sum_{i=1}^{n} W_{in} \xrightarrow{p} 0$. We will use conditional WLLN (Theorem 7) with $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. First note that $\mathbb{E}[W_{in} \mid \mathcal{F}_n] = 0$ by construction. We also check the moment condition (54):

$$\frac{1}{n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n]$$

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

where the latter convergence is by the assumption (Lyap-1). Hence we have that $\frac{1}{n}\sum_{i=1}^{n} W_{in} \mid \mathcal{F}_n \xrightarrow{p,p} 0$ which by Lemma 6 implies $\frac{1}{n}\sum_{i=1}^{n} W_{in} \xrightarrow{p} 0$.

Next we show that $\frac{1}{n}\sum_{i=1}^{n} W_{in}^2 - (\widehat{S_n^{\mathrm{dCRT}}})^2 \xrightarrow{p} 0$. We will use conditional WLLN with $\mathcal{F}_n = \sigma(X, Y, Z)$, observing that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in}^2 \mid \mathcal{F}_n\right] = (\widehat{S_n^{\mathrm{dCRT}}})^2$. Next we verify the moment condition (54):

$$(98) \quad \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n]$$

$$(99) \quad = \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta} \mathbb{E}\left[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z\right] \xrightarrow{p} 0,$$

where the latter convergence is by the assumption (Lyap-1). Hence we have that $\frac{1}{n}\sum_{i=1}^{n} W_{in}^2 - (\widehat{S_n^{\mathrm{dCRT}}})^2 \mid \mathcal{F}_n \xrightarrow{p,p} 0$ which by Lemma 6 implies that $\frac{1}{n}\sum_{i=1}^{n} W_{in}^2 - (\widehat{S_n^{\mathrm{dCRT}}})^2 \xrightarrow{p} 0$.

Combining both of these results we find that

$$(\widehat{S_n^{\mathrm{GCM}}}(\widetilde{X}, X, Y, Z))^2 - (\widehat{S_n^{\mathrm{dCRT}}}(X, Y, Z))^2 \xrightarrow{p} 0.$$

Now using the nondegeneracy condition (NDG1) we can conclude that (97) holds true, as desired. $\qquad\square$

D.2. Proofs of main results.

PROOF OF THEOREM 2. To show the asymptotic equivalence of variance estimates (24) it suffices to show that

$$(100) \quad (\widehat{S_n^{\mathrm{dCRT}}})^2 - s_n^2 \xrightarrow{p} 0; \quad (\widehat{S_n^{\mathrm{GCM}}})^2 - s_n^2 \xrightarrow{p} 0; \quad \inf_n s_n^2 > 0.$$

The first of these statements is given by equation (86) in Lemma 7, the second follows from the proof of Theorem 6 in **?**, and the third is a consequence of assumption (SP2) and conditional independence.

Given the asymptotic equivalence of the variance estimates (24), we can show using Lemma 3 that the GCM test is asymptotically equivalent to the $\widehat{\mathrm{MX}(2)}$ F-test. Indeed, set

$$(101) \quad T_n \equiv T_n^{\mathrm{GCM}} \quad \text{and} \quad C_n \equiv \frac{\widehat{S_n^{\mathrm{dCRT}}}}{\widehat{S_n^{\mathrm{GCM}}}} z_{1-\alpha}$$

Then, ϕ_n^1 is the $\widehat{\mathrm{MX}(2)}$ F-test and ϕ_n^2 is the GCM test. The asymptotic equivalence of the variance estimates (24) then implies the critical value convergence assumption (68) of Lemma 3. The non-accumulation assumption (69) is a consequence of the fact that, under the assumptions of Theorem 2, we have $T_n^{\mathrm{GCM}} \xrightarrow{d} N(0,1)$ (**?**). On the other hand, by Corollary 7 (whose

conclusion holds under the assumptions of Theorem 2), we also know that $\widehat{\text{dCRT}}$ is asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test. Hence, both the GCM test and the $\widehat{\text{dCRT}}$ are asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test, so they are asymptotically equivalent to each other as well.   □

PROOF OF PROPOSITION 1. We treat the two cases separately.

Case 1. We have

$$\frac{1}{n}\sum_{i=1}^{n}((X_i - \widehat{\mu}_{n,x}(Z_i))^2 - \text{Var}_{\mathcal{L}_n}[X_i \mid Z_i])\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{n,x}(Z_i))^2\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \frac{1}{n}\sum_{i=1}^{n}\text{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$= (s_n^2 + o_p(1)) - (s_n^2 + o_p(1))$$

$$= o_p(1),$$

where the third line follows from the convergences (83) and (80) from Lemma 7. This shows the variance consistency property (21).

Case 2. We need to show that

$$\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(X_i)))\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.$$

By the Cauchy-Schwartz inequality, we have

(102)
$$\left| \frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i)))\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \right|$$

$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i))^2\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]}.$$

Given the assumption that $\sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty$ for some $\delta > 0$, Jensen's inequality gives

$$\sup_n \mathbb{E}[\text{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]^{1+\delta/2}] \leq \sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty.$$

Therefore, the weak law of large numbers (Corollary 6) gives

$$\frac{1}{n}\sum_{i=1}^{n}\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \mathbb{E}[\text{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \xrightarrow{p} 0.$$

Furthermore,

$$\mathbb{E}[\text{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \leq \sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}]^{\frac{2}{2+\delta}} < \infty,$$

so

$$\frac{1}{n}\sum_{i=1}^{n}\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] = O_p(1).$$

On the other hand, we know $\text{supp}(\mu_{n,x}(Z_i)) \subseteq \text{Conv}(\text{supp}(\mathcal{L}_n(\boldsymbol{X}))$ for every i and n and by assumption $\widehat{\mu}_{n,x}(\boldsymbol{Z}) \subseteq \text{Conv}(\text{supp}(\mathcal{L}_n(\boldsymbol{X})))$ almost surely. Together with the fact that f is Lipschitz (say with Lipschitz constant L) on $\cup_{n=1}^{\infty}\text{Conv}(\text{supp}(\mathcal{L}_n(\boldsymbol{X})))$, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i)))^2 \text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]$$

$$\leq \frac{L^2}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \equiv L^2(E'_{n,x})^2 = o_p(1).$$

Combining the last two displays with equation (102) gives us the desired result. $\qquad\square$

PROOF OF COROLLARY 2. Define the event

$$A_n \equiv \{\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)\}$$

The conclusion (25) of Theorem 2 implies that $\mathbb{P}_{\mathcal{L}_n}[A_n] \to 0$, so by the assumed contiguity of $\mathcal{L}'_n$ to $\mathcal{L}_n$ we also have $\mathbb{P}_{\mathcal{L}'_n}[A_n] \to 0$. This shows the desired asymptotic equivalence of tests (26). To show the asymptotic equivalence of powers, we derive

$$|\mathbb{E}_{\mathcal{L}'_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) - \phi_n^{\text{GCM}}(X,Y,Z)]|$$

$$\leq \mathbb{E}_{\mathcal{L}'_n}[|\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) - \phi_n^{\text{GCM}}(X,Y,Z)|]$$

$$= \mathbb{E}_{\mathcal{L}'_n}[|\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) - \phi_n^{\text{GCM}}(X,Y,Z)|\mathbb{1}(\phi_n^{\widehat{\text{dCRT}}} \neq \phi_n^{\text{GCM}})]$$

$$\leq \mathbb{E}_{\mathcal{L}'_n}[\mathbb{1}(\phi_n^{\widehat{\text{dCRT}}} \neq \phi_n^{\text{GCM}})]$$

$$= \mathbb{P}_{\mathcal{L}'_n}[\phi_n^{\widehat{\text{dCRT}}} \neq \phi_n^{\text{GCM}}] \to 0.$$

This completes the proof. $\qquad\square$

PROOF OF COROLLARY 3. Fix $\epsilon > 0$, and for each n let $\mathcal{L}_n^* \in \mathscr{L}_n^0 \cap \mathscr{R}_n$ be such that

$$\mathbb{P}_{\mathcal{L}_n^*}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)]$$

$$\geq \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)] - \epsilon.$$

Applying Theorem 2 to the sequence $\mathcal{L}_n^*$ and using the asymptotic Type-I error control of the GCM test, we obtain

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)]$$

$$\leq \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} |\mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{GCM}}(X,Y,Z)]| + \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{GCM}}(X,Y,Z)]$$

$$\leq \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} |\mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{GCM}}(X,Y,Z)]|$$

$$+ \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{GCM}}(X,Y,Z)]$$

$$\leq \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)]$$

$$+ \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{GCM}}(X,Y,Z)]$$

$$\leq \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)] + \alpha$$

$$\leq \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n^*}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)] + \epsilon + \alpha$$

$$= \epsilon + \alpha.$$

Sending $\epsilon \to 0$ gives the desired conclusion. □

PROOF OF THEOREM 4. We have

$$T_n^{\widehat{\mathrm{ndCRT}}}(\widetilde{X},X,Y,Z) \equiv \frac{T_n^{\mathrm{dCRT}}(\widetilde{X},X,Y,Z)}{S_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z)} \equiv \frac{S_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)}{S_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z)} \cdot \frac{T_n^{\mathrm{dCRT}}(\widetilde{X},X,Y,Z)}{S_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)}$$

The first factor converges to 1 in probability (Lemma 9), whereas the second factor converges conditionally on X,Y,Z to $N(0,1)$ (Theorem 1). Putting these two statements together with conditional Slutsky's theorem (Theorem 6), we arrive at the convergence (48). Since the standard normal has continuous CDF we can use Lemma 1 to conclude the convergence of the critical value (49). The equivalence statement (50) follows from the convergence (49) and Lemma 3 applied with $T_n = T_n^{\mathrm{GCM}}$ and $C_n = C_n^{\widehat{\mathrm{ndCRT}}}$. □

PROOF OF COROLLARY 5. The proof of this corollary is directly analogous to that of Corollary 3, so we omit it for the sake of brevity.

□

APPENDIX E: PROOFS FOR SECTION 5

The goal of this section is to prove our main optimality result (Theorem 3) and Corollary 4. The idea of the proof of Theorem 3 is to reduce the problem to a semiparametric testing problem, and then to use existing semiparametric optimality theory. To this end, we first review the relevant semiparametric theory (Section E.1). Then we leverage this theory to prove Theorem 3 (Section E.2) and verify Corollary 4 (Section E.3). Finally, we carry out deferred semiparametric computations (Section E.4).

E.1. Semiparametric preliminaries. Consider a semiparametric model parameterized by

$$(103) \qquad\qquad (\beta, g) \in \mathbb{R} \times \mathcal{H}_g \subseteq \mathbb{R} \times L^2(\nu),$$

where ν is a measure on $\mathbb{R}^p$ and $\mathcal{H}_g \subseteq L^2(\nu)$ is a linear subspace. First, we define a notion of local Type-I error control within the context of the semiparametric model.

DEFINITION 4. Fix a point $g_0 \in \mathcal{H}_g$, and define $\theta_0 \equiv (0, g_0)$. A sequence of tests ϕ_n of $H_0 : \beta = 0$ has asymptotic Type-I error control at θ_0 relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ if, for each submodel $t \mapsto \mathcal{L}_{(0,g_t)}$ with score in $\dot{\mathcal{L}}_{\theta_0}$ along which β is differentiable, we have

$$(104) \qquad\qquad \limsup_{n\to\infty} \mathbb{E}_{\mathcal{L}_{(0,g_{1/\sqrt{n}})}}[\phi_n(W)] \le \alpha.$$

This definition is most similar to that of **?**, except the latter paper does not explicitly use the language of tangent spaces; our definition accommodates Type-I error control over more restricted sets of null distributions reflecting regularity conditions. Next we state a version of the classic semiparametric optimality result:

THEOREM 10 (Theorem 1 in **?**, Theorem 25.44 in **?**, Theorem 18.12 in **?**). *Consider a semiparametric model $\{\mathcal{L}_{\beta,g} : (\beta, g) \in \mathbb{R} \times \mathcal{H}_g\}$ and a point $\theta_0 \equiv (0, g_0)$ for some $g_0 \in \mathcal{H}_g$. Suppose β is differentiable at $\mathcal{L}_{\theta_0}$ relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ with efficient influence function $\widetilde{S}/\widetilde{I}(\theta_0)$, where $\widetilde{S}$ is the efficient score and $\widetilde{I}(\theta_0) > 0$ is the efficient information. For any sequence of tests ϕ_n of $H_0 : \beta = 0$ with asymptotic Type-I error control at θ_0 relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ and any differentiable submodel $\mathcal{L}_t = \mathcal{L}_{(th_\beta, g_t)}$ with score in $\dot{\mathcal{L}}_{\theta_0}$ we have*

$$(105) \qquad\qquad \limsup_{n\to\infty} \mathbb{E}_{\mathcal{L}_{1/\sqrt{n}}}[\phi_n(W)] \le 1 - \Phi(z_{1-\alpha} - h_\beta \cdot \widetilde{I}(\theta_0)^{1/2}).$$

This bound is achieved by the efficient score test $\phi_n^{\mathrm{opt}}(X, Y, Z) \equiv \mathbb{1}(T_n^{\mathrm{opt}}(X, Y, Z) > z_{1-\alpha})$, where

$$(106) \qquad\qquad T_n^{\mathrm{opt}}(X, Y, Z) \equiv \frac{1}{\widetilde{I}(\theta_0)^{1/2} n^{1/2}} \sum_{i=1}^{n} \widetilde{S}(X_i, Y_i, Z_i).$$

In other words,

$$(107) \qquad \begin{aligned} \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_{1/\sqrt{n}}}[\phi_n^{\mathrm{opt}}(W)] &= \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{1/\sqrt{n}}}[T_n^{\mathrm{opt}}(X, Y, Z) > z_{1-\alpha}] \\ &= 1 - \Phi(z_{1-\alpha} - h_\beta \cdot \widetilde{I}(\theta_0)^{1/2}). \end{aligned}$$

This result is like **?**, Theorem 1, except it explicitly deals with tangent spaces. On the other hand, the result is like **?**, Theorem 25.44 or **?**, Theorem 18.12, except it is written in terms of semiparametric models and assumes Type-I error control in the sense of Definition 4 above. By comparison, **??** assume Type-I error control at each point $(0, g)$ for $g \in \mathcal{H}_g$. By inspection of the proof of **?**, Theorem 25.44, only local Type-I error control (Definition 4) is actually needed. In this sense, Theorem 10 can be verified using the same proof as that of **?**, Theorem 25.44, specializing to the case of semiparametric models.

E.2. Proof of Theorem 3. To apply the semiparametric theory from the previous section, the following lemma (proved in Section E.4) identifies the tangent space, the efficient score, and the efficient information at $\mathcal{L}_{\theta_0}$. These results are not novel or surprising; similar results are stated, for example, by **?** in the cases of linear, logistic, and Poisson regressions. Nevertheless, we state and prove Lemma 10 for a self-contained exposition and for precisely tracking the technical assumptions used.

LEMMA 10. [Proof] *In the context of the semiparametric model* (29)*, suppose the following assumptions hold:*

$$(108) \qquad s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]] > 0;$$

$$(109) \qquad \ddot{\psi} = K > 0 \text{ and } \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty \text{ OR } \mathrm{supp}(\boldsymbol{X}, \boldsymbol{Z}) \text{ is compact and } \mathcal{H}_g \subseteq C(\mathbb{R}^p),$$

$$(110) \qquad \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\,\cdot\,] \in \mathcal{H}_g.$$

For each $h = (h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$*, the parametric submodel* $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ *is differentiable in quadratic mean at* $t = 0$ *with score function*

$$(111) \qquad S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \big(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})\big)(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])$$

and satisfies the following local asymptotic normality:

$$(112) \qquad \log \prod_{i=1}^n \frac{\mathrm{d}\mathcal{L}_{\theta_n(h)}}{\mathrm{d}\mathcal{L}_{\theta_0}}(X_i, Y_i, Z_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S(X_i, Y_i, Z_i) - \frac{1}{2}\mathrm{Var}_{\mathcal{L}_{\theta_0}}[S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})] + o_{\mathcal{L}_{\theta_0}}(1).$$

The parameter β *is differentiable at* $\mathcal{L}_{\theta_0}$ *relative to the tangent space*

$$(113) \qquad \dot{\mathcal{L}}_{\theta_0} \equiv \{\big(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})\big)(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) : (h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g\}$$

with efficient score function

$$(114) \qquad \widetilde{S}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = (\boldsymbol{X} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]),$$

efficient information

$$(115) \qquad \widetilde{I}(\theta_0) = s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]],$$

and efficient influence function equal to the ratio of the efficient information and the efficient score.

Note that assumptions (109) and (110) of Lemma 10 are the same as assumptions (36) and (37) of Theorem 3 in the main text; they are restated here for the reader's convenience. Using Lemma 10 in conjunction with Theorem 10, we can prove Theorem 3.

PROOF OF THEOREM 3. Let ϕ_n be a level α test of H_0 as defined in equation (32), and fix $g_0 \in \mathcal{S}$. By assumption (38), $\theta_n(0, h_g) \in \mathcal{R}$ for all $h_g \in \mathcal{H}_g$ for all sufficiently large n. Therefore, ϕ_n also has asymptotic Type-I error control at $\theta_0 \equiv (0, g_0)$ relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ (113) in the sense of Definition 4. Indeed, it suffices to take submodels $t \mapsto \mathcal{L}_{(0, g_t)}$ for $g_t = g_0 + th_g$ and $h_g \in \mathcal{H}_g$, so that $\mathcal{L}_{(0, g_{1/\sqrt{n}})} = \mathcal{L}_{\theta_n(0, h_g)}$. By Lemma 10 (applicable because its first assumption (108) is implied by assumption (35) of Theorem 3 and its last two assumptions are also assumed by Theorem 3), the assumptions of Theorem 10 are met with efficient score $\widetilde{S}$ (114) and efficient information $s^2(\theta_0)$ (115), so taking submodels $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ we find

$$(116) \qquad \limsup_{n \to \infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n(X, Y, Z)] \leq 1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)).$$

On the other hand, because $\mathcal{L}_{\theta_0} \in \mathscr{R}$, it follows that

$$T_n^{\mathrm{GCM}}(X,Y,Z) = \frac{1}{s(\theta_0)\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}_{\mathcal{L}_{\theta_0}}[X_i \mid Z_i])(Y_i - \mathbb{E}_{\mathcal{L}_{\theta_0}}[Y_i \mid Z_i]) + o_{\mathcal{L}_{\theta_0}}(1)$$

$$(117) \qquad\qquad = \frac{1}{\widetilde{I}(\theta_0)^{1/2} n^{1/2}} \sum_{i=1}^{n} \widetilde{S}(X_i, Y_i, Z_i) + o_{\mathcal{L}_{\theta_0}}(1)$$

$$= T_n^{\mathrm{opt}}(X,Y,Z) + o_{\mathcal{L}_{\theta_0}}(1).$$

The first equality follows from the proof of Theorem 6 in **?**, the second follows from the derivations of the efficient score (114) and efficient information (115) in Lemma 10, and the third from equation (106) in Theorem 10. From the local asymptotic normality (112) it follows that $\prod_{i=1}^{n} \mathcal{L}_{\theta_n(h)}$ and $\prod_{i=1}^{n} \mathcal{L}_{\theta_0}$ are contiguous by Le Cam's first lemma (**?**, Example 6.5). It follows that

$$T_n^{\mathrm{GCM}}(X,Y,Z) = T_n^{\mathrm{opt}}(X,Y,Z) + o_{\mathcal{L}_{\theta_n(h)}}(1)$$

We therefore find that

$$1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)) = \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{\theta_n(h)}}[T_n^{\mathrm{opt}} > z_{1-\alpha}] = \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{\theta_n(h)}}[T_n^{\mathrm{GCM}} > z_{1-\alpha}].$$

The first inequality follows from the conclusion (107) of Theorem 10 and the second equality follows from equation (117) and Le Cam's first lemma. Therefore we have shown that for any $h \in (0, \infty) \times \mathcal{H}_g$ and any level α conditional independence test ϕ_n, we have

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

This shows that ϕ_n^{GCM} is LAUMP(g_0) and verifies the claimed asymptotic power (39). Furthermore, since $g_0 \in \mathcal{S}$ was chosen arbitrarily, it follows that ϕ_n^{GCM} is also LAUMP($\mathcal{S}$). This completes the proof. $\qquad\square$

E.3. Proof of Corollary 4. It suffices to verify each of the four assumptions of Theorem 3.

Verification of assumption (35). Note that assumption (SP2) is satisfied because by construction, $X - \mu_x(Z)$ and $Y - \mu_y(Z)$ are independent standard normal random variables for any $\mathcal{L} \in \mathscr{R}$. Next, let

$$(118) \qquad\qquad k(z, z') = \sum_{j=1}^{\infty} \lambda_j e_j(z) e_j(z')$$

be an eigendecomposition of the Sobolev kernel k with eigenfunctions e_j orthonormal with respect to Unif$[0, 1]$. To verify assumption (SP1) given assumption (SP2), it suffices to prove the following statements (**?**, Theorem 11 and Remark 12):

$$\mathrm{Var}_\mathcal{L}[X|Z], \mathrm{Var}_\mathcal{L}[Y|Z] \leq \sigma^2 < \infty \quad \text{almost surely, for all } \mathcal{L} \in \mathscr{R};$$

$$\sup_{\mathcal{L}\in\mathscr{R}} \max(\|\mu_{n,x}\|_{W^{1,2}[0,1]}, \|\mu_{n,y}\|_{W^{1,2}[0,1]}) < \infty;$$

$$\sum_{j=1}^{\infty} \lambda_j < \infty.$$

The first two of these statements follow directly from the construction of $\mathscr{R}$. The third follows from the eigendecomposition of the Sobolev kernel under the uniform measure on $[0, 1]$ (**?**, Example 12.23) with $\lambda_j = (\frac{2}{(2j-1)\pi})^2$.

Verification of assumption (36). Since we are using the normal exponential family, we have $\ddot{\psi} = 1$. Furthermore,

$$(119) \qquad \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] = \mathbb{E}_{\mathcal{L}_{x,z}}[\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2|\boldsymbol{Z}]] = \mathbb{E}_{\mathcal{L}_{x,z}}[\mu_{0x}(\boldsymbol{Z})^2 + 1] < \infty,$$

since $\mu_{0x} \in W^{(1,2)}([0,1]) \subseteq L^2([0,1])$.

Verification of assumption (37). By construction, $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\,\cdot\,] = \mu_{0x} \in W^{1,2}([0,1])$.

Verification of assumption (38). This assumption is a consequence of the openness of the ball $B_{W^{1,2}}(0,C)$ in $W^2([0,1])$.

E.4. Proof of Lemma 10.

Differentiability of parametric submodels. Consider the parametric submodel $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ for some $(h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$, and denote

$$\eta_t(x,z) \equiv xth_\beta + g_0(z) + th_g(z).$$

Letting λ_y be the base measure of the exponential family f_η, we denote $\lambda \equiv \mathcal{L}_{x,z} \times \lambda_y$ and $\mathrm{d}\mathcal{L}_{(th_\beta, g_0 + th_g)}(x,y,z)/\mathrm{d}\lambda$ the density of the parametric model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ with respect to λ. According to **?**, Lemma 7.6, this submodel is differentiable in quadratic mean at $t = 0$ if the map

$$t \mapsto \sqrt{\frac{\mathrm{d}\mathcal{L}_{(th_\beta, g_0 + th_g)}}{\mathrm{d}\lambda}(x,y,z)} = \sqrt{\frac{\mathrm{d}f_{\eta_t}}{\mathrm{d}\lambda_y}(y)} = \exp\left(\frac{y\eta_t - \psi(\eta_t)}{2}\right)$$

is continuously differentiable at $t = 0$ for each $(x,y,z) \in \mathbb{R}^{1+1+p}$ and the elements of the Fisher information matrix are well-defined and continuous at $t = 0$. To show continuous differentiability of the square root density, we compute that

$$\frac{\partial}{\partial t}\sqrt{\frac{\mathrm{d}\mathcal{L}_{(th_\beta, g_0 + th_g)}}{\mathrm{d}\lambda}(x,y,z)} = \exp\left(\frac{y\eta_t - \psi(\eta_t)}{2}\right) \cdot \frac{(y - \dot{\psi}(\eta_t))}{2} \cdot (xh_\beta + h_g(z)).$$

The linearity of η_t in t and the smoothness of ψ imply the continuous differentiability of the above function in t.

Next consider the information matrix

$$I_t \equiv \mathbb{E}_{\mathcal{L}_{(th_\beta, g_0 + th_g)}}\left[\left(\frac{\partial}{\partial t}\log\frac{\mathrm{d}\mathcal{L}_{(th_\beta, g_0 + th_g)}}{\mathrm{d}\lambda}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{L}_{(th_\beta, g_0 + th_g)}}\left[\left(\frac{\partial}{\partial t}(\boldsymbol{Y}\eta_t(\boldsymbol{X}, \boldsymbol{Z}) - \psi(\eta_t(\boldsymbol{X}, \boldsymbol{Z})))\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{L}_{(\beta_t, g_t)}}[(\boldsymbol{Y} - \dot{\psi}(\eta_t(\boldsymbol{X}, \boldsymbol{Z})))^2(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2]$$

$$= \mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2\ddot{\psi}(\eta_t(\boldsymbol{X}, \boldsymbol{Z}))\right].$$

We must show that I_t is well-defined and continuous at $t = 0$. By assumption (109), either $\ddot{\psi} = K > 0$ and $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty$ or $(\boldsymbol{X}, \boldsymbol{Z})$ is compactly supported and $\mathcal{H}_g \subseteq C(\mathbb{R}^p)$. If $\ddot{\psi} = K > 0$ and $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty$, then we have

$$I_0 = C\mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2\right] \leq 2Kh_\beta^2\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] + 2K\mathbb{E}_{\mathcal{L}_{x,z}}[h_g^2(\boldsymbol{Z})]$$

$$= 2Kh_\beta^2\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] + 2K\|h_g\|_{L^2(\nu)}^2 < \infty.$$

Note that, for the sake of this proof, we denote

(120) $\qquad \nu \equiv \mathcal{L}_{x,z}(\boldsymbol{Z}), \quad$ so that $\quad \|h_g\|_{L^2(\nu)}^2 \equiv \mathbb{E}_{\mathcal{L}_{x,z}}[h_g(\boldsymbol{Z})^2] < \infty \quad$ for $h_g \in \mathcal{H}_g.$

Hence, I_0 is well-defined. I_t is also continuous at $t = 0$ because it does not depend on t.

On the other hand, suppose $(\boldsymbol{X}, \boldsymbol{Z})$ is compactly supported and $\mathcal{H}_g \subseteq C(\mathbb{R}^p)$. The quantity inside the expectation defining I_0 is a bounded random variable, because the assumed continuity of h_g implies that this quantity is a continuous function of a random vector $(\boldsymbol{X}, \boldsymbol{Z})$ with compact support. Hence, I_0 is well-defined because it is the expectation of a bounded random variable. To show continuity of I_t, note that by the assumed continuity of h_g and compact support of $(\boldsymbol{X}, \boldsymbol{Z})$ we have $\sup_{t \leq 1} \eta_t(\boldsymbol{X}, \boldsymbol{Z}) \leq B < \infty$ almost surely. Therefore, for $t \leq 1$, we have

$$
\begin{aligned}
|I_t - I_0| &= \left| \mathbb{E}_{\mathcal{L}_{x,z}} \left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2 (\ddot{\psi}(\eta_t) - \ddot{\psi}(\eta_0)) \right] \right| \\
&\leq \mathbb{E}_{\mathcal{L}_{x,z}} \left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2 |\ddot{\psi}(\eta_t) - \ddot{\psi}(\eta_0)| \right] \\
&\leq \mathbb{E}_{\mathcal{L}_{x,z}} \left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2 \sup_{|b| \leq B} |\dddot{\psi}(b)| \cdot |\eta_t - \eta_0| \right] \\
&\leq \sup_{|b| \leq B} |\dddot{\psi}(b)| \cdot \mathbb{E}_{\mathcal{L}_{x,z}} \left[|\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})|^3 \right] \cdot t.
\end{aligned}
$$

We have $\sup_{|b| \leq B} |\dddot{\psi}(b)| < \infty$ because $\dddot{\psi}$ is a continuous function, and $\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})$ almost surely bounded as before. Therefore, we conclude that $|I_t - I_0| \to 0$ as $t \to 0$, so I_t is indeed continuous at 0.

Hence, we conclude by ?, Lemma 7.6 that the parametric submodel $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ is differentiable in quadratic mean at $t = 0$ with score function

$$
\begin{aligned}
S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) &= \frac{\partial}{\partial t} (\boldsymbol{Y}\eta_t(\boldsymbol{X}, \boldsymbol{Z}) - \psi(\eta_t(\boldsymbol{X}, \boldsymbol{Z}))) \Big|_{t=0} \\
&= (\boldsymbol{Y} - \dot{\psi}(\eta_t(\boldsymbol{X}, \boldsymbol{Z})))(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})) \Big|_{t=0} \\
&= (\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]),
\end{aligned}
$$

as claimed (111).

Local asymptotic normality of parametric submodels. The local asymptotic normality of parametric submodels (112) follows from the previously established quadratic mean differentiability (?, Theorem 7.2).

Efficient score and information. For $(h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$, define the score operator

$$
\begin{aligned}
A(h) &\equiv (\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \\
(121) \qquad &= \boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_\beta + (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z}) \\
&\equiv A_\beta(h_\beta) + A_g(h_g).
\end{aligned}
$$

The tangent space $\dot{\mathcal{L}}_{\theta_0}$ (113) can then be expressed as the range of A:

(122) $$\dot{\mathcal{L}}_{\theta_0} = A(\mathbb{R} \times \mathcal{H}_g).$$

As discussed in (?, Section 25.4), the efficient score for β is

(123) $$\tilde{S} = A_\beta - \Pi_{\beta,g} A_\beta,$$

where $\Pi_{\beta,g}$ is the orthogonal projection onto the closure $\overline{A_g(\mathcal{H}_g)}$ of the nuisance tangent space $A_g(\mathcal{H}_g)$ in $L^2(\mathcal{L}_{\theta_0})$. In other words,

$$\text{(124)} \qquad \Pi_{\beta,g}A_\beta = \underset{W \in \overline{A_g(\mathcal{H}_g)}}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})}.$$

To compute this projection, we first claim that the extended operator $A_g : L^2(\nu) \to L^2(\mathcal{L}_{\theta_0})$ is continuous and that $A_g^* A_g$ is continuously invertible. To verify continuity of A_g, note that for $h_g \in L^2(\nu)$ we have

$$
\begin{aligned}
\|A_g(h_g)\|_{L^2(\mathcal{L}_{\theta_0})}^2 &= \mathbb{E}_{\mathcal{L}_{\theta_0}}\left[\left((\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])^2 h_g^2(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\ddot{\psi}(g(\boldsymbol{Z}))h_g^2(\boldsymbol{Z})] \\
&\leq C\mathbb{E}_{\mathcal{L}_{\theta_0}}[h_g^2(\boldsymbol{Z})] \\
&= C\|h_g\|_{L^2(\nu)}^2.
\end{aligned}
$$
(125)

Next, we derive the adjoint operator $A_g^* : L^2(\mathcal{L}_{\theta_0}) \to L^2(\nu)$. For a random variable $W \in L^2(\mathcal{L}_{\theta_0})$, we have

$$
\begin{aligned}
\langle W, A_g h_g \rangle_{L^2(\mathcal{L}_{\theta_0})} &= \mathbb{E}_{\mathcal{L}_{\theta_0}}\left[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z})\right] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \mid \boldsymbol{Z}]h_g(\boldsymbol{Z})] \\
&= \langle \mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \cdot]) \mid \cdot], h_g \rangle_{L^2(\nu)}.
\end{aligned}
$$
(126)

It follows that

$$\text{(127)} \qquad (A_g^* W)(z) = \mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z].$$

Next we derive that

$$
\begin{aligned}
(A_g^* A_g h_\eta)(z) &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[A_g h_\eta(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z})(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \text{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]h_g(z) \\
&= \ddot{\psi}(g(z))h_g(z).
\end{aligned}
$$

The assumption (109) implies that $\ddot{\psi}(g_0(\boldsymbol{Z})) \geq c > 0$ almost surely, since either $\ddot{\psi}$ is a nonzero constant or $g_0(\boldsymbol{Z})$ belongs to a compact set almost surely and therefore $\ddot{\psi}(g_0(\boldsymbol{Z}))$ belongs to the range of a positive continuous function applied to a compact set. From this it follows that $S_g^* S_g$ is continuously invertible. Because A_g is a continuous linear operator with continuously invertible $A_g^* A_g$, it follows that $A_g(L^2(\nu))$ is closed and that $A_g(A_g^* A_g)^{-1}A_g^*$ is the orthogonal projection onto this space. Next let us compute the orthogonal projection of the score A_β onto $A_g(L^2(\nu))$. We have

$$
\begin{aligned}
(A_g^* A_\beta)(z) &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[A_\beta(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z])^2 \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z} = z]\ddot{\psi}(g(z)),
\end{aligned}
$$
(128)

and therefore

$$A_g(A_g^*A_g)^{-1}A_g^*A_\beta = (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])((A_g^*A_g)^{-1}A_g^*A_\beta)(\boldsymbol{Z})$$

$$(129) \qquad = (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}]$$

$$= A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]).$$

Therefore, we have

$$(130) \qquad \underset{W \in A_g(L^2(\nu))}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]).$$

Since $A_g(L^2(\nu))$ is closed, it follows that $\overline{A_g(\mathcal{H}_g)} \subseteq A_g(L^2(\nu))$. Together with the assumption (110) that $\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,] \in \mathcal{H}_g$ and the definition of the effective score as a projection onto $A_g(\mathcal{H}_g)$ (124), we deduce that

$$\Pi_{\beta,g}A_\beta = \underset{W \in \overline{A_g(\mathcal{H}_g)}}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = \underset{W \in A_g(L^2(\nu))}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]).$$

Therefore, the efficient score is

$$\tilde{S} = A_\beta - \Pi_{\beta,g}A_\beta$$

$$(131) \qquad = \boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) - (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}]$$

$$= (\boldsymbol{X} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]),$$

as claimed (114). From there we find that the efficient information is

$$(132) \qquad \widetilde{I}_{\theta_0} = \mathrm{Var}_{\mathcal{L}_{\theta_0}}[\tilde{S}] = \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]] \equiv s^2(\theta_0),$$

also as claimed (115).

Differentiability of β and efficient influence function. By **?**, Lemma 25.25, the differentiability of β at $\mathcal{L}_{\theta_0}$ with respect to the tangent set $\dot{\mathcal{L}}_{\theta_0}$ follows from the quadratic mean differentiability proved above and the assumption that $\widetilde{I}_{\theta_0} = s^2(\theta_0) > 0$ (108). The same lemma gives the efficient influence function as the ratio of the efficient score and the efficient information. This completes the proof.

APPENDIX F: ADDITIONAL MATERIAL RELATED TO SIMULATIONS.

In this section, we present details about existing robustness simulation setups (Section F.1), investigate the trade-off between using lasso and post-lasso (Section F.2) and present the complete simulation results (Section F.3).

F.1. Simulation setup in literature. Here we provide details about the simulation setups considered in **???**.

?. In this paper, the authors consider the double high-dimensional linear model. Suppose $\{Z_i, Y_i\}_{i=1}^n$ is $n = 800$ iid data and $Z_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$ where $p = 800$ and $\boldsymbol{\Sigma}_p$ is chosen to be AR(1) and the autocorrelation is set to be 0.5. Then they consider $Y_i = Z_i^\top \beta + \epsilon$ where β is vector of dimension p and only $s = 50$ is set to be nonzero with magnitude $\nu = 0.175$ and random sign. They consider two ways to set the nonzero components of β: spacing these non-zero coefficients equally or choosing them to be the first 50 coefficients of β. The authors consider, rather than testing conditional independence, the false discovery rate (FDR) of variable selection.

?. In this paper, the authors consider a bit different setting where $Y|Z$ is now a high-dimensional logistic model. The sample size $n = 800$ and $Z_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ where $p = 1500$ and $\boldsymbol{\Sigma}_p$ is chosen to be AR(1) and the autocorrelation is set to be 0.3. After the sampling, the design matrix is centered and every column is normalized to have norm 1. Similarly, only $s = 50$ coordinates of β are set to be nonzero and the sign is random whereas the magnitude is set to be $\nu = 20$. They set a randomly-chosen set of $s = 50$ coefficients of β to be nonzero and consider again the FDR control.

?. In this paper, a similar setting is considered while Z is a data matrix with row $n = 250$ and column $p = 500$ where each row is sampled from $N(0, \boldsymbol{\Sigma}_p)$ and $\boldsymbol{\Sigma}_p$ is an AR(1) matrix with autocorrelation 0.5. However, a crucial difference in this paper is the way to set $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$. As for X, it is generated by a linear predictor $X = Z\gamma + \epsilon$ where γ is a p dimensional vector with first $s = 5$ is nonzero and the other coordinate remains zero and ϵ follows the standard normal distribution. The sign of each coordinate is randomly chosen and the magnitude is set to be $\nu = 0.3$. As for Y, we set $\beta = \gamma$ and $Y = Z\beta + \xi$ where ξ follows standard normal distribution such that ξ is independent of ϵ. We can see that X and Y support on the same subset of Z so that the marginal association between $\mathbf{X}$ and $\mathbf{Y}$ is much larger than that in first two simulation designs.

F.2. Comparing the lasso and post-lasso estimation methods. Compare to the lasso estimation method for $\mu_{n,x}$ and $\mu_{n,y}$, the post-lasso estimation method results in estimates with lower bias but higher variance. This impacts the Type-I error and power of the inferential methods in different ways. For Type-I error, it is only important to have good estimates for $\mathbb{E}[\mathbf{X}|\mathbf{Z}_\mathcal{A}] \equiv \mathbf{Z}_\mathcal{A}^T \beta_\mathcal{A}$ and $\mathbb{E}[\mathbf{Y}|\mathbf{Z}_\mathcal{A}] \equiv \mathbf{Z}_\mathcal{A}^T \gamma_\mathcal{A}$, where $\mathcal{A} \subseteq \{1, \dots, p\}$ denotes the set of variables active in both $\mathbb{E}[\mathbf{X}|\mathbf{Z}]$ and $\mathbb{E}[\mathbf{Y}|\mathbf{Z}]$. Indeed, only the shared active coordinates of $\mathbf{Z}$ act as confounders. On the other hand, for power, it is important to have a good estimate for the entire function $\mathbb{E}[\mathbf{Y}|\mathbf{Z}] = \mathbf{Z}^T \gamma$ (**?**). Therefore, we examine the mean-squared estimation error for $\mathbf{Z}_\mathcal{A}^T \beta_\mathcal{A}$ and $\mathbf{Z}_\mathcal{A}^T \gamma_\mathcal{A}$ in one of our null simulation settings as well as the mean-squared estimation error for $\mathbf{Z}^T \beta$ and $\mathbf{Z}^T \gamma$ in one of our alternative simulation settings (Figure 5). We find that the post-lasso does a better job estimating the shared active coefficients in the null setting, so the reduced bias in estimating these shared coefficients outweighs the increased variance. On the other hand, the lasso does a better job estimating the entire set of coefficients, so in this case the increased variance outweighs the reduced bias. This explains why the post-lasso-based methods have improved Type-I error control but worse power than the lasso-based methods.

F.3. Additional simulation results. Figures 6-13 present the complete simulation results across the null and alternative, Gaussian and binary, and supervised and unsupervised settings. All the standard errors are less than 0.026.
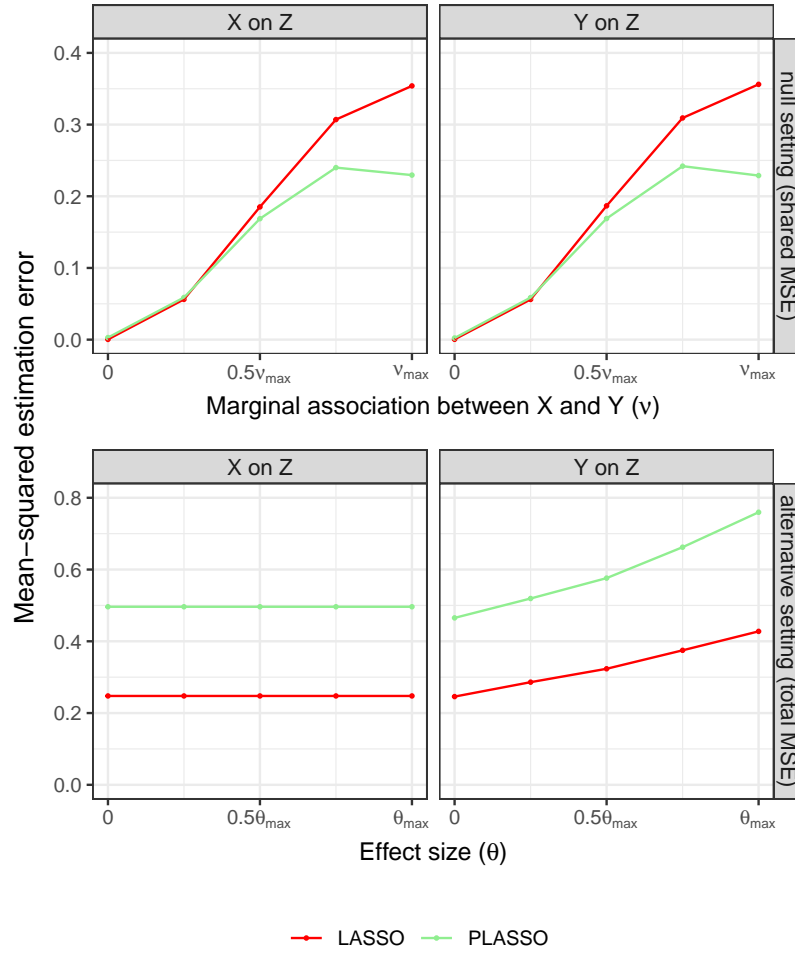
Fig 5: MSE on shared variables and total variables: first column displays the MSE of lasso and post-lasso on shared active variables $Z_{\mathcal{A}}$ and the second column displays the MSE of lasso and post-lasso on total variables. All the experiments are carried out with GCM statistic with $n = 100, d = 400, s = 5, \rho = 0.4$. All the standard errors are less than 0.014.
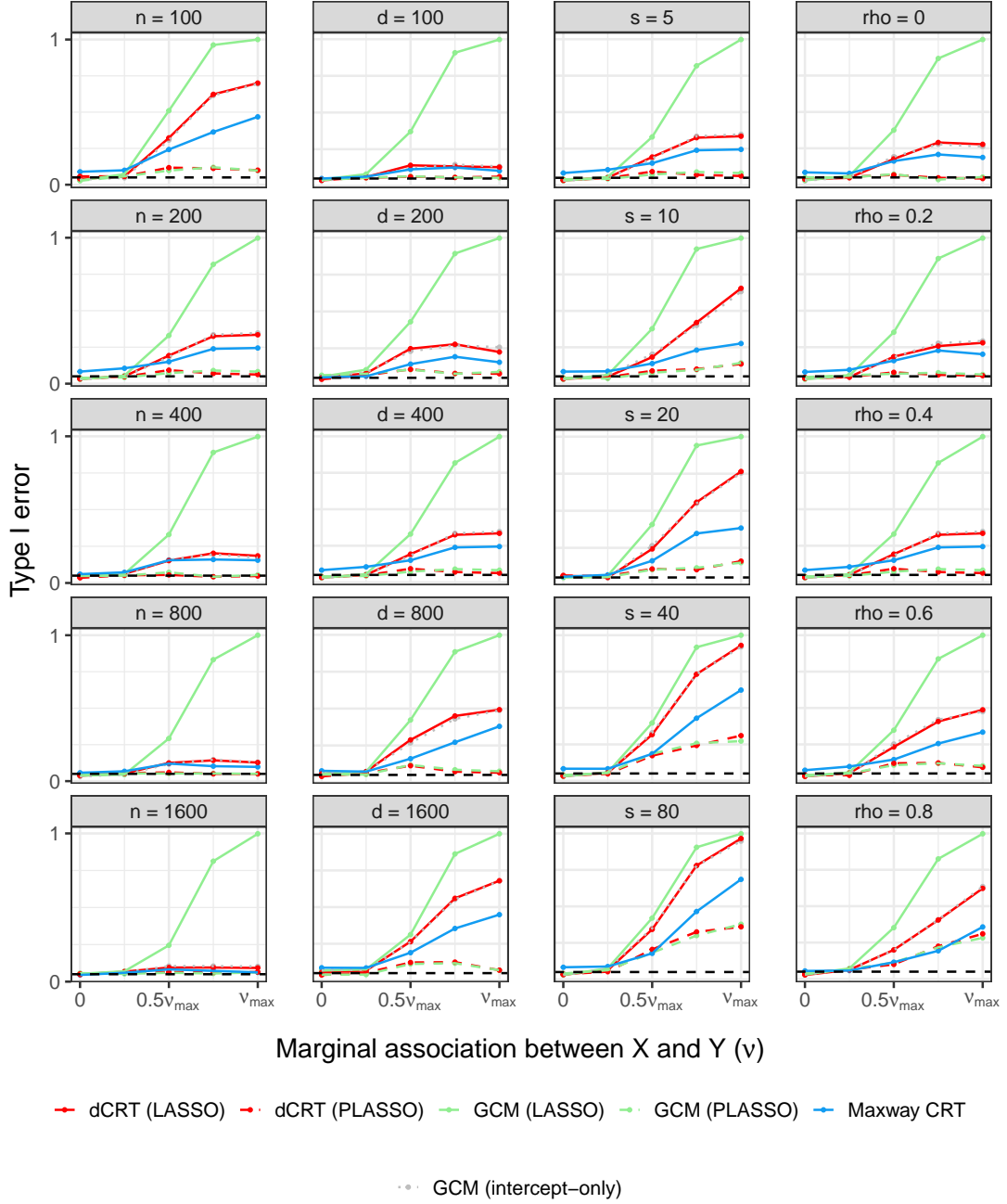
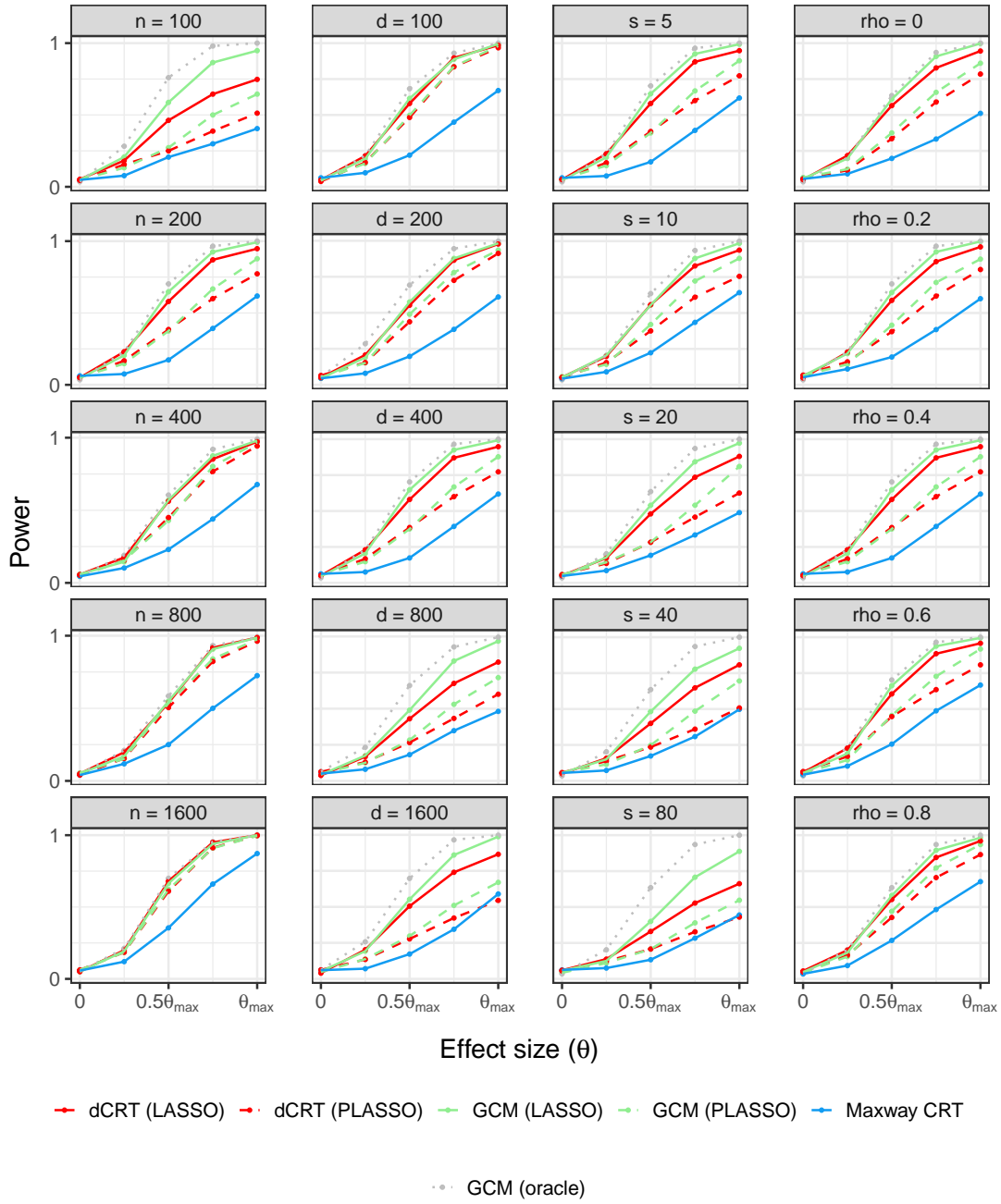Fig 6: Type-I error in the Gaussian supervised setting.
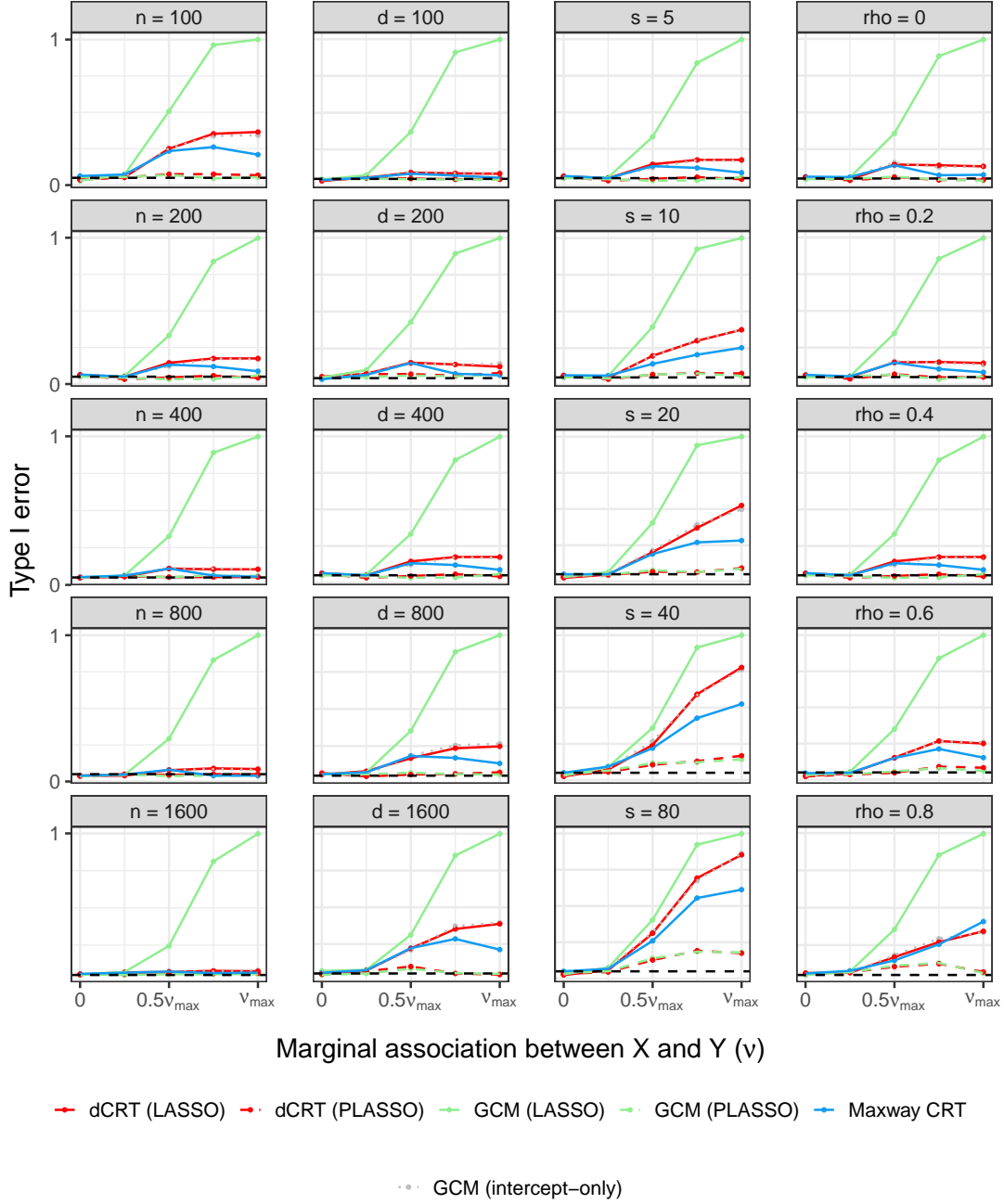
Fig 7: Power in the Gaussian supervised setting.

Fig 8: Type-I error in the Gaussian semi-supervised setting.
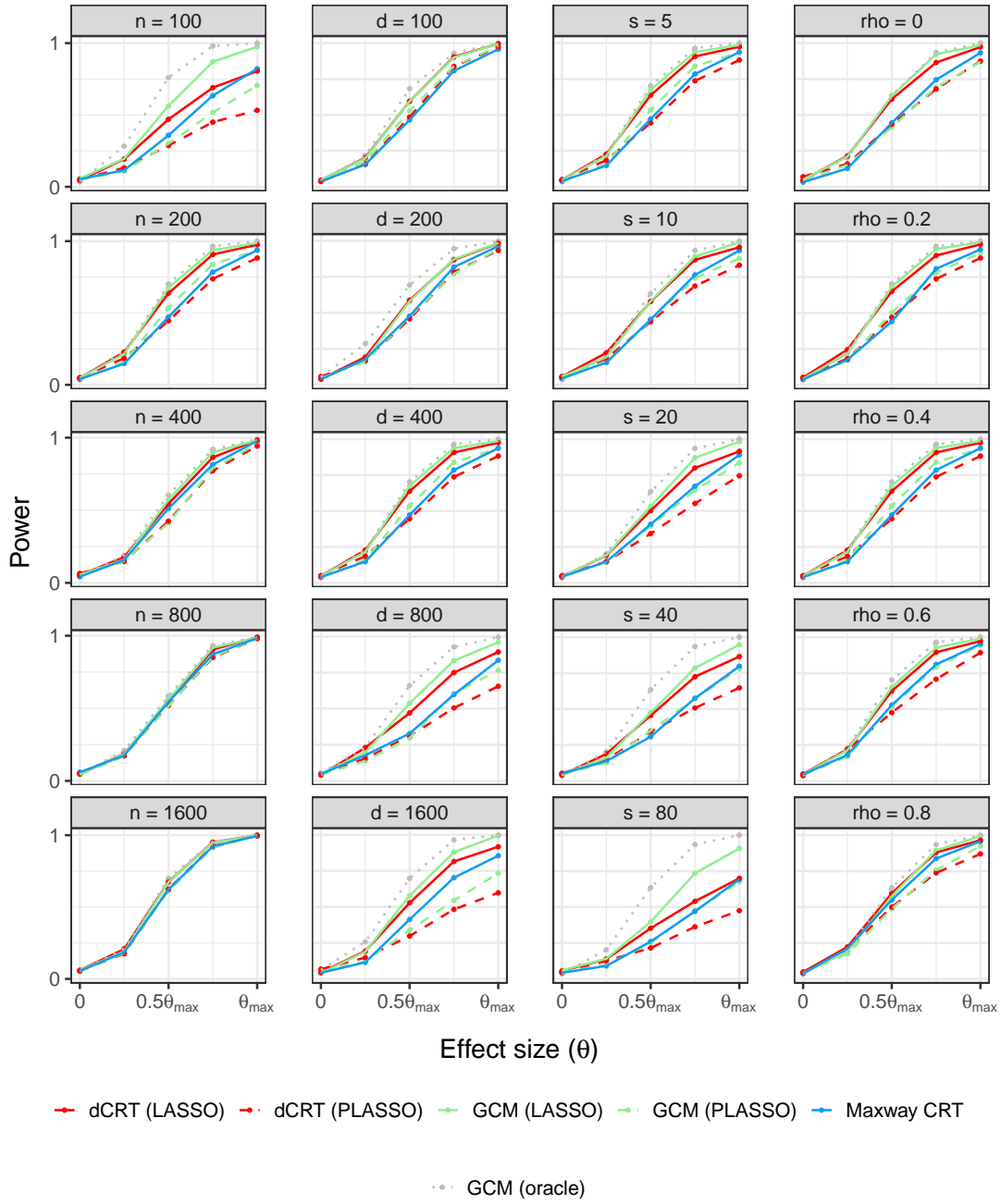
Fig 9: Power in the Gaussian semi-supervised setting.
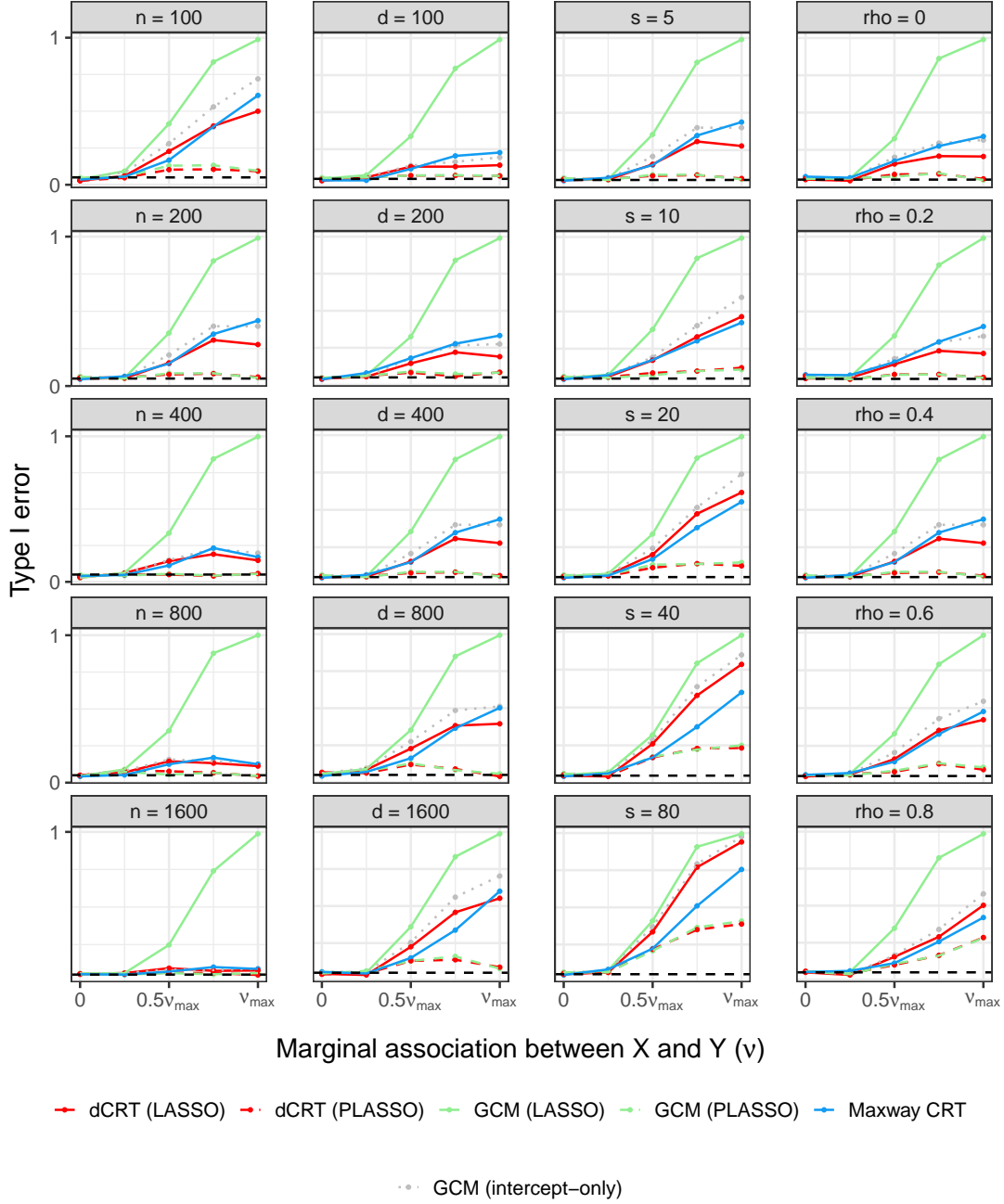
Fig 10: Type-I error in the binary supervised setting.
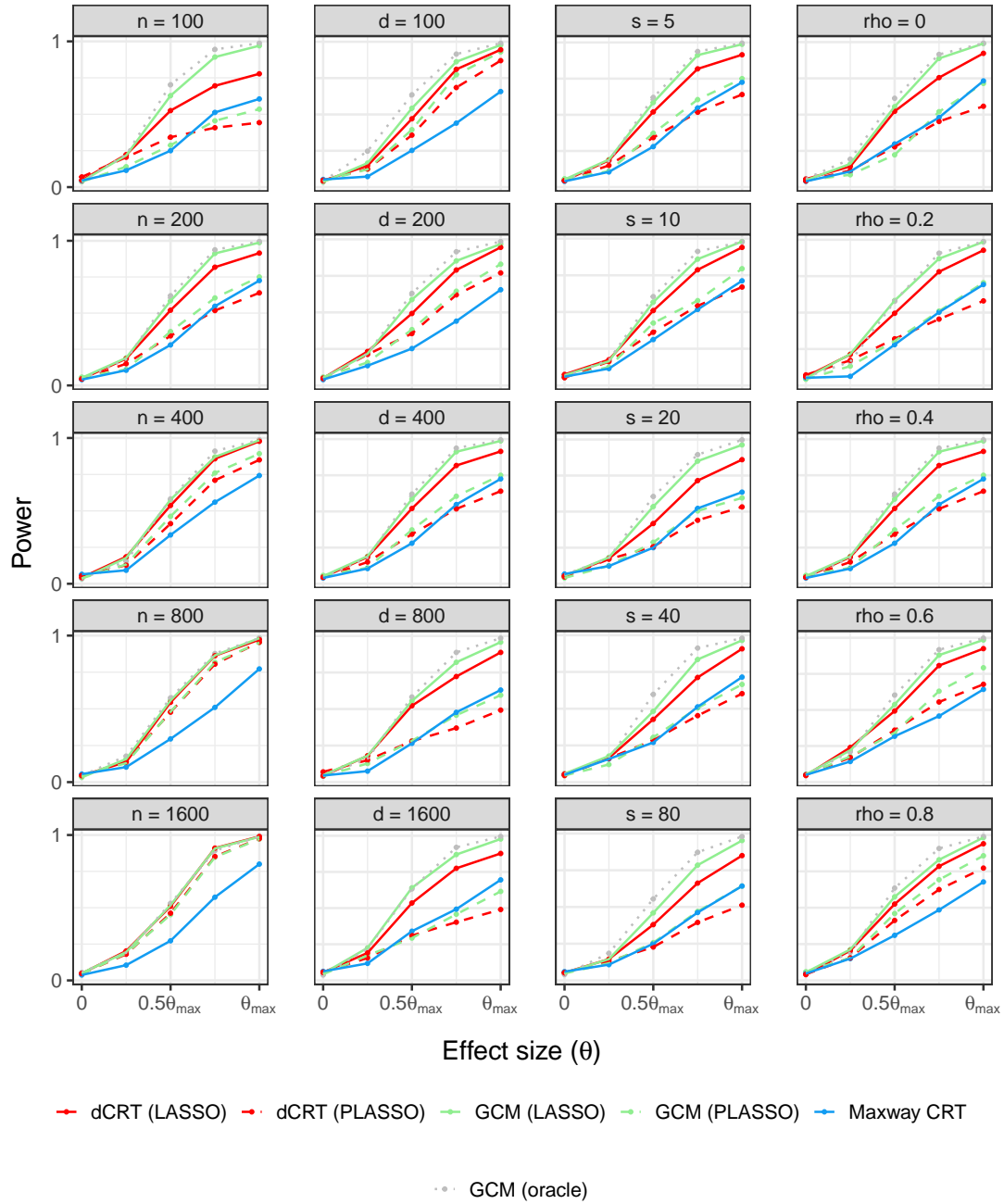
Fig 11: Power in the binary supervised setting.
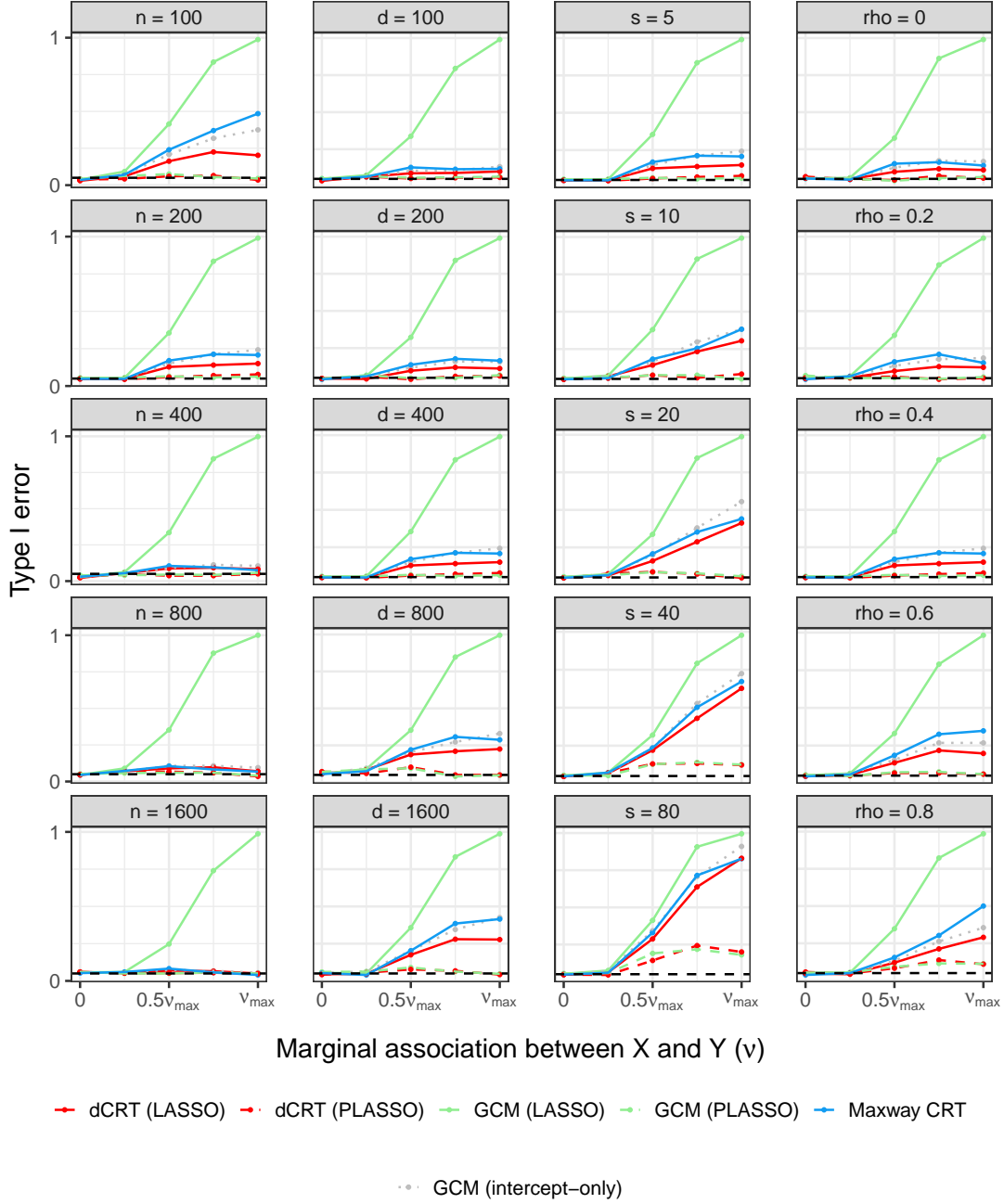
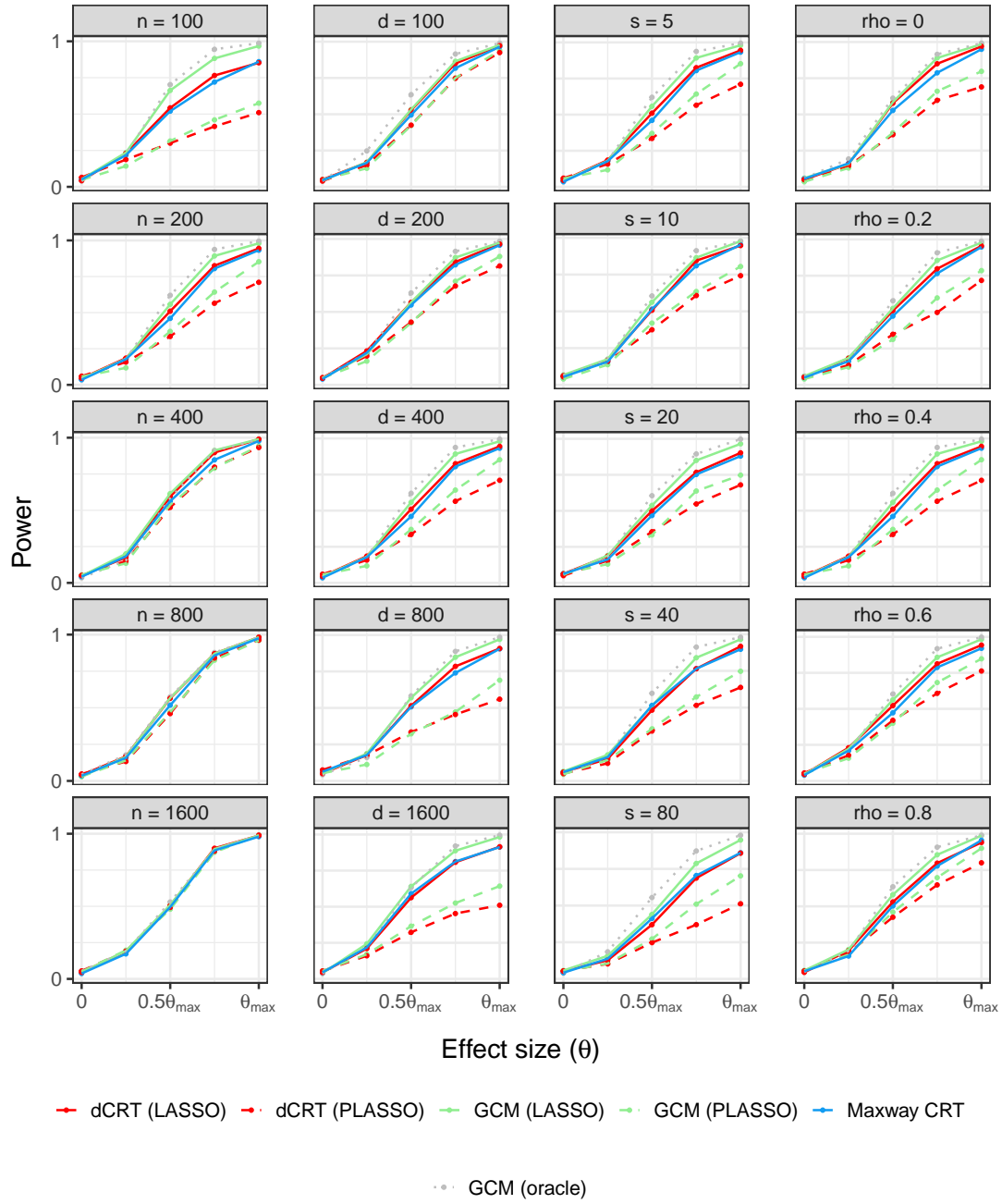Fig 12: Type-I error in the binary semi-supervised setting.

Fig 13: Power in the binary semi-supervised setting.

APPENDIX G: PROOFS OF CONDITIONAL CONVERGENCE RESULTS

In this section, we present the proofs of the conditional convergence results from Appendix B. We proceed by stating and proving necessary lemmas in Section G.1 and then proving the convergence results themselves in Section G.2.

G.1. Auxiliary lemmas.   First we state a few known results for the reader's convenience.

LEMMA 11 (**?**, Theorem 2.3.2).   *A sequence of random variables W_n converges to a limit W in probability if and only if every subsequence of W_n has a further subsequence that converges to W almost surely.*

LEMMA 12 (Conditional Markov inequality, **?**, Theorem 10.17).   *Let W be a random variable and let $\mathcal{F}$ be a σ-algebra. If for some $q > 0$ we have $\mathbb{E}[|W|^q] < \infty$, then for any ϵ we have*

$$\mathbb{P}(|W| \geq \epsilon | \mathcal{F}) \leq \frac{\mathbb{E}[|W|^q | \mathcal{F}]}{\epsilon^q} \quad \text{almost surely.}$$

LEMMA 13 (Conditional Hölder inequality, **?**, Theorem 6.60).   *Let W_1 and W_2 be random variables and let $\mathcal{F}$ be a σ-algebra. If for some $q_1, q_2 \in (1, \infty)$ with $\frac{1}{q_1} + \frac{1}{q_2} = 1$ we have $\mathbb{E}[|W_1|^{q_1}], \mathbb{E}[|W_2|^{q_2}] < \infty$, then*

$$\mathbb{E}[|W_1 W_2| \mid \mathcal{F}] \leq (\mathbb{E}[|W_1|^{q_1} \mid \mathcal{F}])^{1/q_1} (\mathbb{E}[|W_2|^{q_2} \mid \mathcal{F}])^{1/q_2} \quad \text{almost surely.}$$

LEMMA 14 (**?**, Lemma 11.2.1).   *Suppose $W_n \xrightarrow{d} W$. If for given $\alpha \in (0, 1)$ the CDF of W is continuous and strictly increasing at $\mathbb{Q}_\alpha[W]$, then*

$$\mathbb{Q}_\alpha[W_n] \to \mathbb{Q}_\alpha[W].$$

Next we establish that, without of loss of generality, all random variables, σ-algebras, and conditional expectations in a triangular array may be viewed as being defined on a common probability space.

LEMMA 15 (Embedding into a single probability space).   *Consider a sequence of probability spaces $\{(\mathbb{P}_n, \Omega_n, \mathcal{G}_n), n \geq 1\}$. For each n, let $\{W_{i,n}\}_{i \geq 1}$ be a collection of integrable random variables defined on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ and let $\mathcal{F}_n \subseteq \mathcal{G}_n$ be a σ-algebra. Then there exists a single probability space $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$, random variables $\{\widetilde{W}_{i,n}\}_{i,n \geq 1}$ on $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$, and σ-fields $\widetilde{\mathcal{F}}_n \subseteq \widetilde{\mathcal{G}}$ for $n \geq 1$, such that for each n, the joint distribution of $(\{W_{i,n}\}_{i \geq 1}, \{\mathbb{E}[W_{i,n} \mid \mathcal{F}_n]\}_{i \geq 1})$ on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ coincides with that of $(\{\widetilde{W}_{i,n}\}_{i \geq 1}, \{\mathbb{E}[\widetilde{W}_{i,n} \mid \widetilde{\mathcal{F}}_n]\}_{i \geq 1})$ on $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$.*

PROOF.   Define the Cartesian product $\widetilde{\Omega} \equiv \prod_{n=1}^{\infty} \Omega_n$, the σ-algebra $\widetilde{\mathcal{G}}$ generated by measurable cylinders $\prod_{n=1}^{\infty} A_n$ for $A_n \in \mathcal{G}_n$ and $A_n = \mathcal{G}_n$ for all but finitely many n, and the infinite product measure $\widetilde{\mathbb{P}}$ on the measurable space $(\widetilde{\Omega}, \widetilde{\mathcal{G}})$ (**?**). On this probability space, define σ-algebras

(133) $$\widetilde{\mathcal{F}}_n \equiv \{\mathcal{G}_1 \times \cdots \times \mathcal{G}_{n-1} \times A_n \times \mathcal{G}_{n+1} \times \cdots : A_n \in \mathcal{F}_n\}$$

and random variables

(134) $$\widetilde{W}_{in}(\omega) \equiv W_{in}(\omega_n)$$

for each $i, n \geq 1$. Next we claim that for each $i, n \geq 1$, the random variable

$$(135) \qquad \mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n](\omega) \equiv \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n) \quad \text{for each } \omega \in \widetilde{\Omega}$$

is in fact a version of the conditional expectation $\mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n]$. Indeed, it suffices to check that for each $A \equiv \mathcal{G}_1 \times \cdots \times \mathcal{G}_{n-1} \times A_n \times \mathcal{G}_{n+1} \times \cdots \in \widetilde{\mathcal{F}}_n$ we have

(136)

$$\int_A \mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n](\omega)d\widetilde{\mathbb{P}}(\omega) \equiv \int_A \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n)d\widetilde{\mathbb{P}}(\omega)$$

$$= \int_{\prod_{n' \neq n} \Omega_{n'}} \int_{A_n} \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n)d\widetilde{\mathbb{P}}_n(\omega_n)d \prod_{n' \neq n} \mathbb{P}_{n'}(\omega_{n'})$$

$$\equiv \int_{\prod_{n' \neq n} \Omega_{n'}} \int_{A_n} W_{in}(\omega_n)d\widetilde{\mathbb{P}}_n(\omega_n)d \prod_{n' \neq n} \mathbb{P}_{n'}(\omega_{n'})$$

$$= \int_A W_{in}(\omega_n)d\widetilde{\mathbb{P}}(\omega)$$

$$\equiv \int_A \widetilde{W}_{in}(\omega)d\widetilde{\mathbb{P}}(\omega).$$

From the ω-wise embeddings (134) and (135), it is easy to verify the claimed equality between the joint distributions on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ and $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$. $\qquad \square$

Finally, we state a conditional version of the truncated weak law of large numbers:

LEMMA 16. *For each n, let $W_{in}, 1 \leq i \leq n$ be a set of random variables independent conditionally on $\mathcal{F}_n$. Let $b_n > 0$ with $b_n \to \infty$ and let $\bar{W}_{in} = W_{in}\mathbb{1}(|W_{in}| \leq b_n)$. Suppose that as $n \to \infty$ we have*

1. $\sum_{i=1}^n \mathbb{P}[|W_{in}| > b_n|\mathcal{F}_n] \xrightarrow{p} 0$ *and*
2. $b_n^{-2} \sum_{i=1}^n \mathbb{E}[\bar{W}_{in}^2|\mathcal{F}_n] \xrightarrow{p} 0$.

If we set $S_n \equiv \sum_{i=1}^n W_{in}$ and $a_n \equiv \sum_{i=1}^n \mathbb{E}[\bar{W}_{in}]$ then

$$\frac{S_n - a_n}{b_n} \mid \mathcal{F}_n \xrightarrow{p,p} 0.$$

PROOF. Let $\bar{S}_n \equiv \sum_{i=1}^n \bar{W}_{in}$. We first write

$$\mathbb{P}\left[\left|\frac{S_n - a_n}{b_n}\right| > \epsilon \Big| \mathcal{F}_n\right] \leq \mathbb{P}\left[S_n \neq \bar{S}_n|\mathcal{F}_n\right] + \mathbb{P}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon \Big| \mathcal{F}_n\right].$$

To estimate the first term, we note that

$$\mathbb{P}[S_n \neq \bar{S}_n|\mathcal{F}_n] \leq \mathbb{P}[\cup_{i=1}^n \{\bar{W}_{in} \neq W_{in}\}|\mathcal{F}_n] \leq \sum_{i=1}^n \mathbb{P}(|W_{in}| > b_n|\mathcal{F}_n) \xrightarrow{p} 0$$

by the first assumption. For the second term, we note that conditional Markov's inequality (Lemma 12), $a_n = \mathbb{E}[\bar{S}_n|\mathcal{F}_n]$ implies that

$$\mathbb{P}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon|\mathcal{F}_n\right] \leq \epsilon^{-2}\mathbb{E}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right|^2 \Big| \mathcal{F}_n\right]$$

$$= \epsilon^{-2} b_n^{-2} \mathrm{Var}[\bar{S}_n | \mathcal{F}_n]$$

$$= (b_n \epsilon)^{-2} \sum_{i=1}^{n} \mathrm{Var}[\bar{W}_{in} | \mathcal{F}_n]$$

$$\leq (b_n \epsilon)^{-2} \sum_{i=1}^{n} \mathbb{E}\left[\bar{X}_{in}^2 | \mathcal{F}_n\right] \xrightarrow{p} 0,$$

where the convergence in the last line is given by the second assumption. This completes the proof. $\square$

G.2. Proofs of conditional convergence results.

PROOF OF THEOREM 5. This proof generalizes the argument of (**?**, Lemma 2.11) to allow for conditioning. Fix $\epsilon > 0$, and choose an integer $k \geq 2/\epsilon$. Because the CDF of W is continuous, it follows that

$$\mathbb{P}[W \leq \mathbb{Q}_{i/k}[W]] = i/k \quad \text{for each } 0 \leq i \leq k.$$

Fix $t \in \mathbb{R}$, and suppose $\mathbb{Q}_{\frac{i-1}{k}}[W] \leq t \leq \mathbb{Q}_{\frac{i}{k}}[W]$. It follows that

$$\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i-1}{k}}[W] \mid \mathcal{F}_n] - \frac{i}{k} \leq \mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t] \leq \mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \frac{i-1}{k}.$$

Therefore, for all $t \in \mathbb{R}$, we have

$$|\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| \leq \sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{i/k}[W] \mid \mathcal{F}_n] - \frac{i}{k}\right| + \frac{1}{k}$$

$$= \sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| + \frac{1}{k},$$

so that

$$\sup_{t \in \mathbb{R}} |\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| \leq \sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| + \frac{1}{k}.$$

By assumption, we have

$$(137) \qquad \mathbb{P}\left[\left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k(k+1)}\right] \to 0.$$

Therefore,

$$\mathbb{P}\left[\sup_{t \in \mathbb{R}} |\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| > \epsilon\right]$$

$$\leq \mathbb{P}\left[\sup_{t \in \mathbb{R}} |\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| > \frac{2}{k}\right]$$

$$\leq \mathbb{P}\left[\sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k}\right]$$

$$\leq \sum_{i=0}^{k} \mathbb{P}\left[\left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k(k+1)}\right]$$

$$\to 0.$$

This completes the proof. $\square$

PROOF OF THEOREM 6. Fix $t \in \mathbb{R}$. Letting $F(t') \equiv \mathbb{P}[W \leq t']$ be the CDF of W, Theorem 5 gives

$$\left| \mathbb{P}\left[W_n \leq \frac{t - b_n}{a_n} \mid \mathcal{F}_n \right] - F\left(\frac{t - b_n}{a_n} \right) \right| \leq \sup_{t' \in \mathbb{R}} |\mathbb{P}[W_n \leq t' | \mathcal{F}_n] - F(t')| \xrightarrow{p} 0.$$

By the continuous mapping theorem, we have $F\left(\frac{t - b_n}{a_n} \right) \xrightarrow{p} F(t)$, so that

$$\mathbb{P}\left[W_n \leq \frac{t - b_n}{a_n} \mid \mathcal{F}_n \right] \xrightarrow{p} F(t).$$

Noting that $\mathbb{P}[a_n \leq 0 | \mathcal{F}_n]$ is a sequence of nonnegative random variables whose expectations converge to zero, it follows that $\mathbb{P}[a_n \leq 0 | \mathcal{F}_n] \xrightarrow{p} 0$ and so

$$\mathbb{P}\left[a_n W_n + b_n \leq t \mid \mathcal{F}_n \right] = \mathbb{P}\left[a_n W_n + b_n \leq t, a_n > 0 \mid \mathcal{F}_n \right] + o_p(1)$$

$$= \mathbb{P}\left[W_n \leq \frac{t - b_n}{a_n}, a_n > 0 \mid \mathcal{F}_n \right] + o_p(1)$$

$$= \mathbb{P}\left[W_n \leq \frac{t - b_n}{a_n} \mid \mathcal{F}_n \right] + o_p(1)$$

$$\xrightarrow{p} F(t),$$

as desired. $\square$

PROOF OF THEOREM 7. We apply Lemma 16 with $b_n = n$. We first verify the first assumption in Lemma 16 by conditional Markov's inequality (Lemma 12):

$$\sum_{i=1}^n \mathbb{P}(|W_{in}| > n | \mathcal{F}_n) \leq \sum_{i=1}^n \frac{\mathbb{E}(|W_{in}|^{1+\delta} | \mathcal{F}_n)}{n^{1+\delta}} \xrightarrow{p} 0.$$

For the second condition, we have

$$\frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[W_{in}^2 \mathbb{1}(|W_{in}| \leq n) | \mathcal{F}_n] \leq \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[|W_{in}|^{1+\delta} n^{1-\delta} \mathbb{1}(|W_{in}| \leq n)]$$

$$= \frac{1}{n^{1+\delta}} \sum_{i=1}^n \mathbb{E}[|W_{in}|^{1+\delta} \mathbb{1}(|W_{in}| \leq n) | \mathcal{F}_n]$$

$$\leq \frac{1}{n^{1+\delta}} \sum_{i=1}^n \mathbb{E}[|W_{in}|^{1+\delta} | \mathcal{F}_n] \xrightarrow{p} 0.$$

Therefore, Lemma 16 yields

$$\frac{1}{n} \sum_{i=1}^n (W_{in} - \mathbb{E}[W_{in} \mathbb{1}(|W_{in}| \leq n) \mid \mathcal{F}_n]) \,\Bigg|\, \mathcal{F}_n \xrightarrow{p,p} 0.$$

By conditional Slutsky (Theorem 6), it now suffices to show that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[W_{in} \mathbb{1}(|W_{in}| > n) \mid \mathcal{F}_n \right] \xrightarrow{p} 0.$$

To see this, applying conditional Markov's and Hölder's inequalities (Lemmas 12 and 13, respectively) we obtain

$$
\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[W_{in} \mathbb{1}(|W_{in}| > n)|\mathcal{F}_n\right] \right| &\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[|W_{in}| \mathbb{1}(|W_{in}| > n)|\mathcal{F}_n\right] \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \{\mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n]\}^{1/(1+\delta)} \{\mathbb{P}[|W_{in}| > n|\mathcal{F}_n]\}^{\delta/(1+\delta)} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \{\mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n]\}^{1/(1+\delta)} \left\{ \frac{\mathbb{E}(|W_{in}|^{1+\delta}|\mathcal{F}_n)}{n^{1+\delta}} \right\}^{\delta/(1+\delta)} \\
&= \frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n] \\
&\xrightarrow{p} 0,
\end{aligned}
$$

where the last convergence is by assumption. Finally, we verify that the condition (56) is sufficient for the conditional WLLN assumption (54) by noting that it implies

$$
\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] \leq \frac{\sup_{1 \leq i \leq n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] n}{n^{1+\delta}} = \frac{\sup_{1 \leq i \leq n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n]}{n^{\delta}} \xrightarrow{p} 0.
$$

This completes the proof. $\qquad\square$

PROOF OF THEOREM 8. Without loss of generality, we assume $\mathbb{E}[W_{in}|\mathcal{F}_n] = 0$ and that all random variables and σ-algebras are defined on a common probability space $(\mathbb{P}, \Omega, \mathcal{G})$ (Lemma 15). Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra on $\mathbb{R}^n$. Let κ_n be a regular conditional distribution of $(W_{1n}, \ldots, W_{nn})$ given $\mathcal{F}_n$ (**?**, Theorem 8.37), i.e. a function $\kappa_n : \Omega \times \mathcal{B}(\mathbb{R}^n) \to [0, \infty]$ such that $\omega \mapsto \kappa_n(\omega, B)$ is measurable for each $B \in \mathcal{B}(\mathbb{R}^n)$, $B \mapsto \kappa_n(\omega, B)$ is a σ-finite measure on $\mathbb{R}^n$ for each $\omega \in \Omega$, and

$$
\kappa_n(\omega, B) = \mathbb{P}[(W_{1n}, \ldots, W_{nn}) \in B|\mathcal{F}_n](\omega), \quad \text{for almost all } \omega \in \Omega \text{ and all } B \in \mathcal{B}(\mathbb{R}^n).
$$

For each n and each $\omega \in \Omega$, let $(\widetilde{W}_{1n}(\omega), \ldots, \widetilde{W}_{nn}(\omega))$ be a draw from the measure $\kappa_n(\omega, \cdot)$. By **?**, Theorem 8.38, we have for each n that

$$
\left(\sum_{i=1}^{n} \text{Var}[\widetilde{W}_{in}(\omega)] \right)^{-(2+\delta)/2} \sum_{i=1}^{n} \mathbb{E}[|\widetilde{W}_{in}(\omega)|^{2+\delta}] \overset{a.s.}{=} \frac{1}{S_n^{2+\delta}(\omega)} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n](\omega).
$$

Now, let $\{n_k\}_{k \geq 1}$ be a subsequence of $\mathbb{N}$. By the conditional Lyapunov assumption (61) and Lemma 11, there is a further subsequence n_{k_j} such that

$$
(138) \qquad \frac{1}{S_{n_{k_j}}^{2+\delta}} \sum_{i=1}^{n_{k_j}} \mathbb{E}[|W_{in_{k_j}}|^{2+\delta} \mid \mathcal{F}_{n_{k_j}}] \overset{a.s.}{\to} 0.
$$

Hence, it follows that
$$
(139)
$$
$$
\left(\sum_{i=1}^{n_{k_j}} \text{Var}[\widetilde{W}_{in_{k_j}}(\omega)] \right)^{-(2+\delta)/2} \sum_{i=1}^{n_{k_j}} \mathbb{E}[|\widetilde{W}_{in_{k_j}}(\omega)|^{2+\delta}] \to 0 \quad \text{for almost every } \omega \in \Omega.
$$

Applying the usual Lyapunov CLT to the triangular array $\{\widetilde{W}_{in_{k_j}}(\omega)\}_{i,n_{k_j}}$, we find that

(140) $\quad \left(\sum_{i=1}^{n_{k_j}} \text{Var}[\widetilde{W}_{in_{k_j}}(\omega)]\right)^{-1/2} \sum_{i=1}^{n_{k_j}} \widetilde{W}_{in_{k_j}}(\omega) \xrightarrow{d} N(0,1) \quad \text{for almost every } \omega \in \Omega,$

and therefore that, for each $t \in \mathbb{R}$, we have

(141)
$$\mathbb{P}\left[\left(\sum_{i=1}^{n_{k_j}} \text{Var}[\widetilde{W}_{in_{k_j}}(\omega)]\right)^{-1/2} \sum_{i=1}^{n_{k_j}} \widetilde{W}_{in_{k_j}}(\omega) \leq t\right] \to \Phi(t) \quad \text{for almost every } \omega \in \Omega.$$

Using **?**, Theorem 8.38 again, it follows that for each $t \in \mathbb{R}$, we have

(142)
$$\mathbb{P}\left[\frac{1}{S_{n_{k_j}}} \sum_{i=1}^{n_{k_j}} W_{in_{k_j}} \leq t \mid \mathcal{F}_{n_{k_j}}\right] \xrightarrow{a.s.} \Phi(t).$$

Applying Lemma 11, it follows that

(143)
$$\mathbb{P}\left[\frac{1}{S_n} \sum_{i=1}^{n} W_{in} \leq t \mid \mathcal{F}_n\right] \xrightarrow{p} \Phi(t),$$

as desired. $\qquad\square$

PROOF OF LEMMA 1. Without loss of generality, we assume that all random variables and σ-algebras are defined on a common probability space $(\mathbb{P}, \Omega, \mathcal{G})$ (Lemma 15). Let $\mathcal{B}(\mathbb{R})$ be the Borel σ-algebra on $\mathbb{R}$. Let κ_n be a regular conditional distribution of W_n given $\mathcal{F}_n$ (**?**, Theorem 8.29), i.e. a function $\kappa_n : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, \infty]$ such that $\omega \mapsto \kappa_n(\omega, B)$ is measurable for each $B \in \mathcal{B}(\mathbb{R})$, $B \mapsto \kappa_n(\omega, B)$ is a σ-finite measure on $\mathbb{R}^n$ for each $\omega \in \Omega$, and

$$\kappa_n(\omega, B) = \mathbb{P}[W_n \in B | \mathcal{F}_n](\omega), \quad \text{for almost all } \omega \in \Omega \text{ and all } B \in \mathcal{B}(\mathbb{R}).$$

Now, let $\{n_k\}_{k \geq 1}$ be a subsequence of $\mathbb{N}$. By conditional Polya's theorem (Theorem 5), we have

(144)
$$\sup_{t \in \mathbb{R}} |\mathbb{P}[W_n \leq t | \mathcal{F}_n] - \mathbb{P}[W \leq t]| \xrightarrow{p} 0.$$

Hence, by Lemma 11 there is a further subsequence n_{k_j} such that

(145)
$$\sup_{t \in \mathbb{R}} |\mathbb{P}[W_{n_{k_j}} \leq t | \mathcal{F}_{n_{k_j}}](\omega) - \mathbb{P}(W \leq t)| \to 0, \text{ for almost all } \omega \in \Omega.$$

It follows that

(146)
$$\sup_{t \in \mathbb{R}} |\kappa_{n_{k_j}}(\omega, (-\infty, t]) - \mathbb{P}(W \leq t)| \to 0, \text{ for almost all } \omega \in \Omega,$$

i.e.

(147)
$$\kappa_{n_{k_j}}(\omega, \cdot) \xrightarrow{d} W \text{ for almost all } \omega \in \Omega.$$

Hence, by Lemma 14, it follows that

(148)
$$\mathbb{Q}_\alpha[\kappa_{n_{k_j}}(\omega, \cdot)] \to \mathbb{Q}_\alpha[W] \text{ for almost all } \omega \in \Omega,$$

and therefore

(149)
$$\mathbb{Q}_\alpha[W_{n_{k_j}} | \mathcal{F}_{n_{k_j}}] \xrightarrow{a.s.} \mathbb{Q}_\alpha[W].$$

Applying Lemma 11 again, we conclude that

(150)
$$\mathbb{Q}_\alpha[W_n | \mathcal{F}_n] \xrightarrow{p} \mathbb{Q}_\alpha[W],$$

as desired. $\qquad\square$

APPENDIX H: FINITE-SAMPLE IMPROVEMENT USING $\widehat{\text{ndCRT}}$ OVER GCM

We proved that the $\widehat{\text{dCRT}}$ and GCM test are asymptotically equivalent, both under the null and under local alternatives (Theorem 2 and Corollary 2). Despite this asymptotic equivalence, the two methods can behave quite differently in finite samples. In particular, the resampling-based $\widehat{\text{dCRT}}$ method can have better finite-sample performance than the asymptotic GCM test. In this appendix, we present a simple numerical example to illustrate this phenomenon, focusing on the Type-I error of the two methods. In order to isolate the effect of the resampling rather than of the test statistic, we perform the comparison using the GCM test statistic for both methods. Therefore, we consider the $\widehat{\text{ndCRT}}$ variant (Appendix A) rather than the $\widehat{\text{dCRT}}$ itself.

Consider the null data-generating model

$$\boldsymbol{Y} \sim \text{Poisson}(\mu_y(\boldsymbol{Z})),\ \boldsymbol{X} \sim \text{Bern}(\mu_x(\boldsymbol{Z}))$$

where $\mu_y(\boldsymbol{Z}) = \exp(\beta_0 + \boldsymbol{Z}\beta_1)$ and $\mu_x(\boldsymbol{Z}) = \text{expit}(\gamma_0 + \boldsymbol{Z}\gamma_1)$ and $\boldsymbol{Z} \sim N(0, 1)$. In other words, $\boldsymbol{Y}|\boldsymbol{Z}$ is a Poisson regression model and $\boldsymbol{X}|\boldsymbol{Z}$ is a logistic regression model. We consider the following set of parameters:

$$\gamma_0 = -4,\ \gamma_1 = 1,\ \beta_0 = -3,\ \beta_1 = 1,$$

chosen to make both X and Y sparse. Such a setting occurs commonly in single-cell genomics, and is one where we would expect the convergence of the central limit theorem to be slower. In this simulation, we take the sample size $n = 1000$ and employ the usual generalized linear model estimates for $\mu_x(\cdot)$ and $\mu_y(\cdot)$. We set $M = 1000$ when applying $\widehat{\text{ndCRT}}$ and the simulation is repeated 1000 times.

Figure 14 shows the results of the simulation. The top-left panel compares the p-values from the two methods across 1000 realizations of the data, showing that the $\widehat{\text{ndCRT}}$ produces well-calibrated p-values while the GCM test produces inflated p-values. Thresholding these p-values as nominal level $\alpha = 0.05$, we find that the $\widehat{\text{ndCRT}}$ method controls the Type-I error at the nominal level, while the GCM test does not (top-right panel). The reason for the miscalibration of the GCM test in this setting is that the null distribution of the test statistic has not yet converged to normality (bottom-left panel). By contrast, the resampling distribution of the GCM statistic (employed by the $\widehat{\text{ndCRT}}$ to obtain the critical value) aligns well with its null distribution (bottom-right panel), explaining the improved calibration of the $\widehat{\text{ndCRT}}$ method. Finally, the bottom-right panel shows the empirical CDF of the test statistics, which illustrates the difference in the null distributions of the two methods.
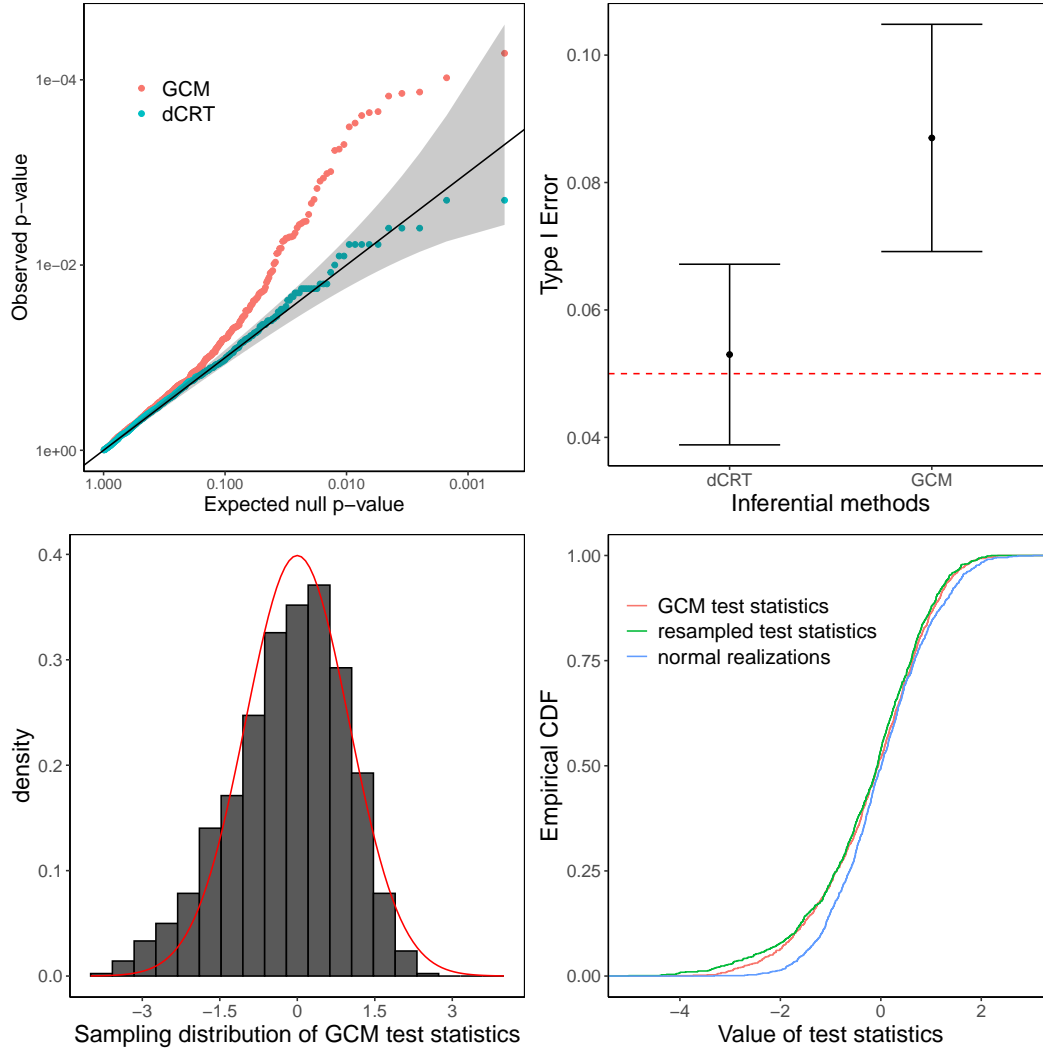
Fig 14: Comparing the calibration of the GCM and $\widehat{\text{ndCRT}}$ methods in a sparse regime. Top left: QQ-plots of the p-values produced by the two methods in a null simulation setting. Top right: Type-I error of the two methods at $\alpha = 0.05$. Bottom left: The true null distribution of the GCM statistic (gray histogram), as well as the standard normal density (red curve). Bottom right: Comparisons of the empirical CDFs of the GCM test statistics, the $\widehat{\text{ndCRT}}$ resamples, and the standard normal distribution.

APPENDIX I: EFFECT EVALUATION OF ASYMMETRY NATURE IN $\widehat{\text{dCRT}}$

It is intuitively true that with similar sample sizes, estimating a Gaussian linear model is easier than a Poisson linear model. With such intuition, we would expect that different re-sampling strategies will result in different Type-I error control when one considers Gaussian-Poisson model, i.e. $\boldsymbol{Y}|\boldsymbol{Z}$ is a Poisson model and $\boldsymbol{X}|\boldsymbol{Z}$ is a Gaussian model. This is because with fewer sample size it is more difficult to estimate the Poisson model due to the exponential link function. In this section, we provide a set of simulation study to verify such intuition. We consider the following model:

$$\boldsymbol{X}|\boldsymbol{Z} \sim N(\boldsymbol{Z}^\top \gamma, 1), \ \boldsymbol{Y}|\boldsymbol{Z} \sim \text{Poisson}(\exp(-4 + \boldsymbol{Z}^\top \beta)), \ \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_p).$$

Consider the simulation setting when $n = 200$ and $p = 10, 12, \ldots, 18$. The fitting methods applied are GLM for both $\boldsymbol{X}|\boldsymbol{Z}$ and $\boldsymbol{Y}|\boldsymbol{Z}$. Also the unnormalized statistic, $T_n^{\widehat{\text{dCRT}}}$, and normalized statistic $T_n^{\widehat{\text{ndCRT}}}$ are both considered. The simulation is independently repeated 5000 times. The Type-I error results are presented in Figure 15. The most phenomenal observations are as follows:

(a) P-values obtaining from resampling from Gaussian model, together with p-value based on GCM, are mcuh more aligned with the significance level;

(b) When using normalized statistics, significantly more type-I error appears when resamples are drawn from Poisson model;

(c) When using unnormalized statistics, resampling from Poisson model will result in very conservative p-values.

It turns out that the performance of unnormalized test statistics observed in (a) and (c) can be explained by the second plot in Figure 16. Notice the unnormalized resampling test statistics are well aligned with sampling test statistics when resmapling from $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$. However, there is an extremely high variation in the resampled test statistics when resampling from $\widehat{\mathcal{L}}_n(\boldsymbol{Y}|\boldsymbol{Z})$. This makes the corresponding p-values very conservative given the comparison between test statistics and resmapling test statistics.

As for normalized statistics, the resampling variance can make huge impact on the Type-I error control. Specifically, the third plot compares the empirical variance for $\widehat{\text{dCRT}}$ statistic when resampling from $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z}), \widehat{\mathcal{L}}_n(\boldsymbol{Y}|\boldsymbol{Z})$, which are defined respectively as,

$$\frac{1}{n-1} \sum_{i=1}^n \left((\widetilde{X}_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)) - \frac{1}{n} \sum_{i=1}^n (\widetilde{X}_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)) \right)^2$$

and

$$\frac{1}{n-1} \sum_{i=1}^n \left((X_i - \hat{\mu}_{n,x}(Z_i))(\widetilde{Y}_i - \hat{\mu}_{n,y}(Z_i)) - \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{n,x}(Z_i))(\widetilde{Y}_i - \hat{\mu}_{n,y}(Z_i)) \right)^2.$$

Similarly, we compute the empirical variance for the unnormalized test statistic (so-called sampling empirical variance in the third plot in Figure 16)

$$\frac{1}{n-1} \sum_{i=1}^n \left((X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)) - \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)) \right)^2,$$

which serves as the normalization in normalized statistic (up to factor $(n-1)/n$). Clearly, the resampling variance when resampling from $Y|Z$ is drastically large than resampling variance when resampling from $X|Z$. Even though the resampling distribution for unnormalized test statistics in second plot in Figure 16 also displays very large variation, this variation has
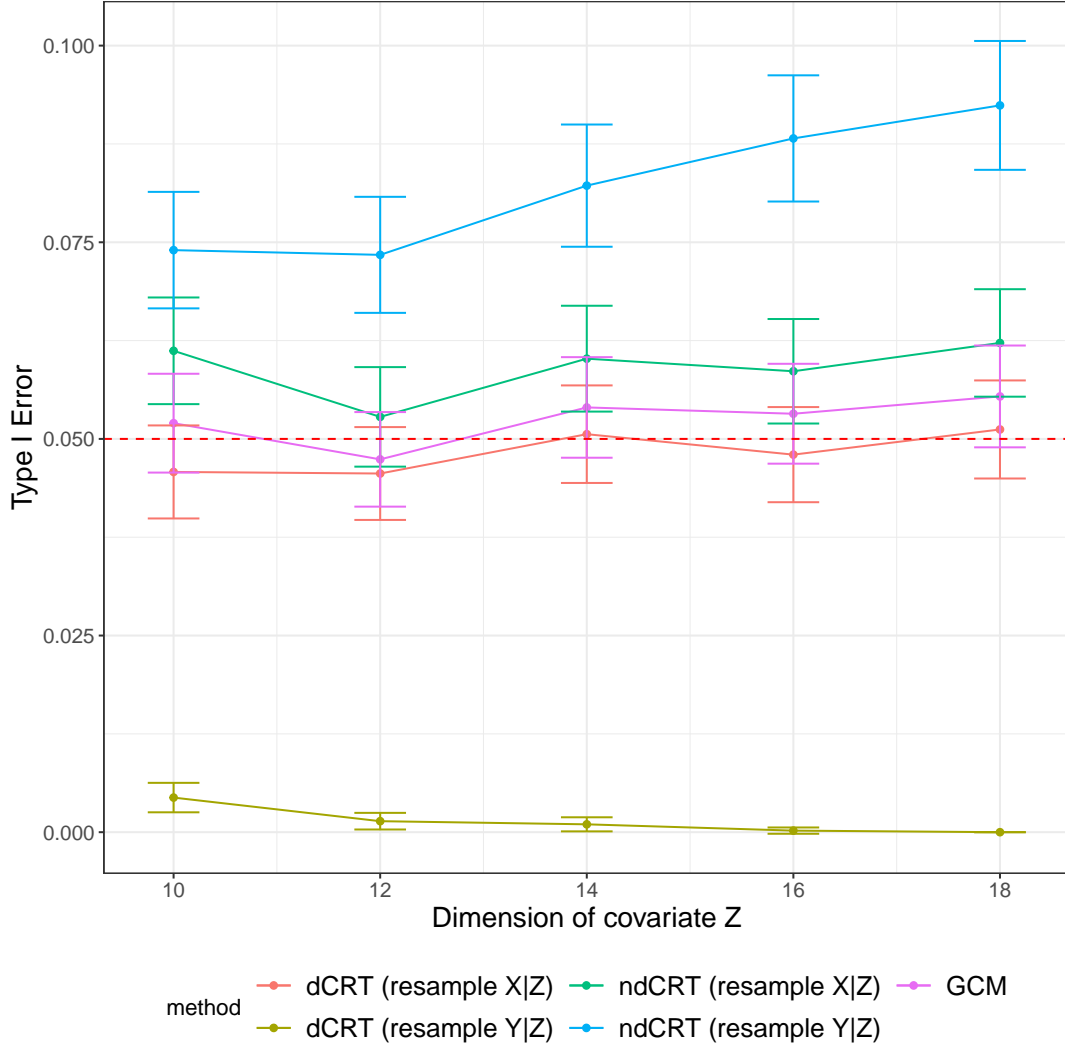
Fig 15: Type-I error comparison when $Y|Z$ is a Poisson model.

been compensated by the normalization. This causes the observation (b) the absurd multi-modality phenomenon in the frist plot of Figure 16. Moreover, the first plot shows that the range in both histograms of the resampled test statistics from $\widehat{\mathcal{L}}_n(Y|Z)$ and $\widehat{\mathcal{L}}_n(X|Z)$ are narrower than the normalied test statistic while the former is more severe. This explains the wild Type-I error behvaior of $\widehat{\mathrm{ndCRT}}$ when resampling from $\widehat{\mathcal{L}}_n(Y|Z)$ (see blue and yellow lines in Figure 15).

We believe that the Type-I error inflation/deflation when resampling from $\widehat{\mathcal{L}}_n(Y|Z)$ is due to the amplifing effect of the exponential funciton in the conditional mean $\mu_y(\cdot)$. In partic-ular, the exponential function will produce extremely large resamples when the conditional mean happens to be large at some given covariate value. The large variation will make the unnormalized test very conservative and the normalized statistic anti-consevative. Therefore, the asymmetry nature in $\widehat{\mathrm{dCRT}}$ can be a double-edged sword depending on the underlying data generating mechanism.
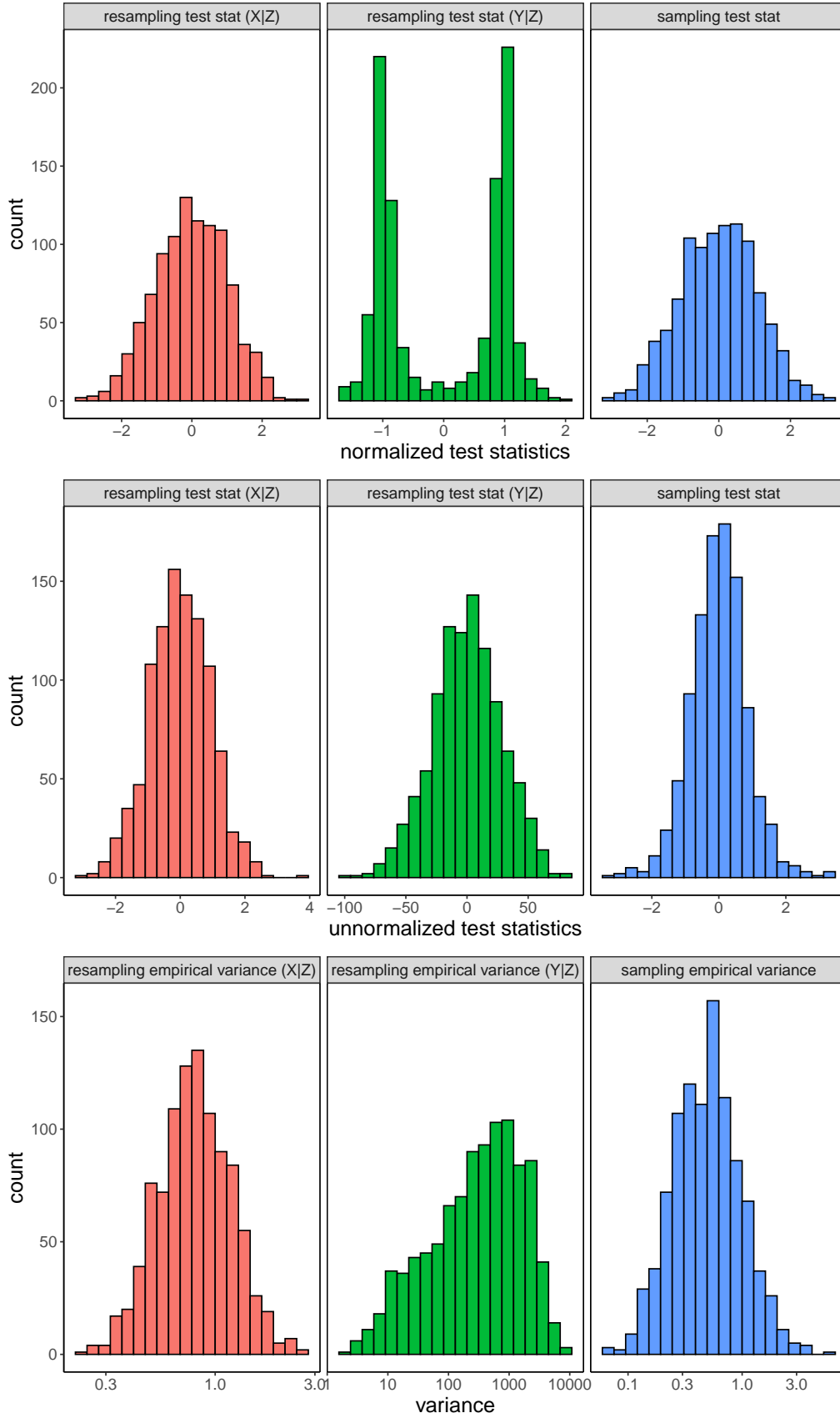
Fig 16: Histograms for resampled test statistics, sampling test statistics and empirical variances.