On the power of conditional independence testing under model-X

Eugene Katsevich[1] and Aaditya Ramdas[1,2]

ekatsevi@wharton.upenn.edu, aramdas@stat.cmu.edu
Department of Statistics and Data Science, University of Pennsylvania[1]
Department of Statistics and Data Science, Carnegie Mellon University[2]
Machine Learning Department, Carnegie Mellon University[2]

July 20, 2021

Abstract

For testing conditional independence (CI) of a response Y and a predictor X given covariates Z, the recently introduced model-X (MX) framework has been the subject of active methodological research, especially in the context of MX knockoffs and their successful application to genome-wide association studies. In this paper, we study the power of MX CI tests, yielding quantitative explanations for empirically observed phenomena and novel insights to guide the design of MX methodology. We show that any valid MX CI test must also be valid conditionally on Y, Z; this conditioning allows us to reformulate the problem as testing a point null hypothesis involving the conditional distribution of X. The Neyman-Pearson lemma then implies that the conditional randomization test (CRT) based on a likelihood statistic is the most powerful MX CI test against a point alternative. We also obtain a related optimality result for MX knockoffs. We show that the CRT resampling distribution of an appropriately normalized test statistic converges to a standard normal distribution under the MX assumption, leading to an asymptotically equivalent test based on explicit critical values (and therefore no need for resampling) with uniform asymptotic type-I error control. In fact, this test is valid under the assumption that *only the first two moments of X given Z are known*, a significant relaxation of MX. Finally, we derive expressions for the power of this test (and that of the asymptotically equivalent CRT) against local semiparametric alternatives, explicitly capturing the prediction error of the underlying learning algorithm.

1 Introduction

1.1 Conditional independence testing and the MX assumption

Given a predictor $\boldsymbol{X} \in \mathbb{R}^d$, response $\boldsymbol{Y} \in \mathbb{R}^r$, and covariate vector $\boldsymbol{Z} \in \mathbb{R}^p$ drawn from a joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}$, consider testing the hypothesis of conditional independence (CI),

$$H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z} \quad \text{versus} \quad H_1 : \boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}, \tag{1}$$

using n data points

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

This fundamental problem—determining whether a predictor is associated with a response after controlling for a set of covariates—is ubiquitous across the natural and social sciences. To keep an example in mind throughout the paper, consider $\boldsymbol{Y} \in \mathbb{R}^1$ cholesterol level, $\boldsymbol{X} \in \{0, 1, 2\}^{10}$ the genotypes of an individual at 10 adjacent polymorphic sites, and $\boldsymbol{Z} \in \{0, 1, 2\}^{500{,}000}$ the genotypes of the individual at other polymorphic sites across the genome. Such data (X, Y, Z) would be collected in a genome-wide association study (GWAS), with the goal of testing for association between the 10 polymorphic sites of interest and cholesterol while controlling for the other polymorphic sites (1). CI testing is also connected to causal inference: with appropriate unconfoundedness assumptions, Fisher's sharp null hypothesis of no effect of a (potentially non-binary) treatment $\boldsymbol{X}$ on an outcome $\boldsymbol{Y}$ implies conditional independence. While we do not work in a causal framework, we draw inspiration from connections to causal inference throughout.

As formalized by Shah and Peters [1], the problem (1) is fundamentally impossible without assumptions on the distribution $\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, in which case no asymptotically uniformly valid test of this hypothesis can have nontrivial power against any alternative. In special cases, the problem is more tractable, for example if $\boldsymbol{Z}$ has discrete support, or if we were willing to make (semi)parametric assumptions on the form of $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ (henceforth "model-$Y|X$"). We will not be making such assumptions in this work. Instead, we follow the lead of Candes et al. [2], who proposed to avoid assumptions on $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, but assume that we have access to $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$:[1]

$$\textit{model-X (MX) assumption} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}} \text{ for a known } f^*_{\boldsymbol{X}|\boldsymbol{Z}}. \tag{3}$$

Candes et al argue that while both model-$Y|X$ and MX are strong assumptions—especially when p, d are large—in certain cases much more is known about $\boldsymbol{X}|\boldsymbol{Z}$ than about $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$. In the aforementioned GWAS example, $\boldsymbol{X}|\boldsymbol{Z}$ reflects the joint distribution of genotypes at SNPs across the genome, which is well described by hidden Markov models from population genetics [3]. On the other hand, the distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ represents the genetic basis of a complex trait, about which much less is known. In the context of (stratified)

[1] Candes et al actually require that the full joint distribution $\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z})$ is known, but this is because they also test for conditional associations between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. We focus only on the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ and therefore require a weaker assumption.

randomized experiment, the distribution $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is the propensity function [4] (the analog of the propensity score for non-binary treatments [5]) and is experimentally controlled. In general causal inference contexts, the MX assumption can be viewed as the assumption that the propensity function is known.

1.2 MX methodology and open questions

Testing CI hypotheses in the MX framework has been the subject of active methodological research. The most popular methodology is MX knockoffs [2]. This method is based on the idea of constructing synthetic negative controls (knockoffs) for each predictor variable in a rigorous way that is based on the MX assumption; see Section 5.1 for a brief overview. Rapid progress has been made on the construction of knockoffs in various cases [3, 6, 7, 8] and on the application of this methodology to GWAS [3, 9]. The conditional randomization test (CRT) [2], initially less popular than knockoffs due to its computational cost, is receiving renewed attention as computationally efficient variants are proposed, such as the holdout randomization test (HRT) [10], the digital twin test [11], and the distilled CRT (dCRT) [12]. We describe this methodology next.

We start with any test statistic $T(X, Y, Z)$ measuring the association between $\boldsymbol{X}$ and $\boldsymbol{Y}$, given $\boldsymbol{Z}$. Usually, this statistic involves learning some estimate $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ based on machine learning, e.g. the magnitude of the fitted coefficient for $\boldsymbol{X}$ (assuming $\dim(\boldsymbol{X}) = 1$) in a cross-validated lasso [13] of Y on X and Z [2] . To calculate the distribution of T under the null hypothesis (1), first define a matrix $\widetilde{X} \in \mathbb{R}^{n\times d}$ where the ith row $\widetilde{X}_i$ is a sample from $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z} = Z_i)$. In other words, for each sample i, we resample X_i based on its distribution conditional on the observed covariate values Z_i in that sample. We then use these resamples to build a null distribution $T(\widetilde{X}, Y, Z)$, from which we extract the upper quantile

$$C_\alpha(Y, Z) \equiv Q_{1-\alpha}[T(\widetilde{X}, Y, Z)|Y, Z], \tag{4}$$

where the randomness is over the resampling distribution $\widetilde{X}|Y, Z$. Then, the CRT rejects if the original test statistic exceeds this quantile:

$$\phi_T^{\mathrm{CRT}}(X, Y, Z) \equiv \begin{cases} 1, & \text{if } T(X, Y, Z) > C_\alpha(Y, Z); \\ \gamma, & \text{if } T(X, Y, Z) = C_\alpha(Y, Z); \\ 0, & \text{if } T(X, Y, Z) < C_\alpha(Y, Z). \end{cases} \tag{5}$$

In order to accommodate discreteness, the CRT makes a randomized decision when $T(X, Y, Z) = C_\alpha(Y, Z)$ so that the size of the test is exactly α. In practice, the threshold $C_\alpha(Y, Z)$ is approximated by computing $T(\widetilde{X}^b, Y, Z)$ for a large number B of Monte Carlo resamples $\widetilde{X}^b \sim X|Z$. For the sake of clarity, in this paper we consider only the "infinite-B" version of the CRT as defined by equations (4) and (5). In the causal inference setting, the CRT can be viewed as a variant of Fisher's exact test for randomized experiments that incorporates strata of covariates [14, 15], basing inference on rerandomizing the treatment to the units.

The conditional independence testing problem under MX has benefited from a variety of methodological innovations, but there are still many open theoretical questions about this problem, including the following three:

Q1. Are there "optimal" test statistics for MX methods, in any sense?

Q2. To what extent can the MX assumption be weakened?

Q3. What is the precise connection between the performance of the machine learning (ML) step and the power of the resulting MX method?

In this paper, we shed light on these questions. We summarize our findings next.

1.3 Our contributions

We find that the CRT is a natural setting to analyze the MX CI problem; it is simpler to analyze than MX knockoffs and is applicable for testing a single conditional independence hypothesis. For these reasons, we focus mainly on the CRT in the present paper. We obtain the following (partial) answers to the questions posed above.

A1: MX CI is a conditional inference problem and a CRT is most powerful against point alternatives. While the composite alternative of the CI problem (1) suggests that we cannot expect to find a uniformly most powerful test, we may still ask what is the most powerful test against a point alternative. We show that any level α test must also have level α conditionally on (Y, Z), which allows us to reduce the composite null to a point null. We can therefore apply the Neyman-Pearson lemma to show (Section 2) that the optimal test against a point alternative $\mathcal{L}$ with $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ is the CRT based on the likelihood test statistic:

$$T^{\text{opt}}(X; Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{6}$$

The same statistic yields the most powerful one-bit p-values for MX knockoffs (Section 5). Since the model for $Y|X, Z$ is unknown, this result provides our first theoretical indication of the usefulness of ML models to learn this distribution (Q3). A3 below gives a more quantitative answer to Q3.

A2: The MX assumption can be drastically weakened while retaining asymptotic Type-I error control. Huang and Janson [8] recently showed that finite-sample type-I error control is possible under only the assumption that the model for $(\boldsymbol{X}, \boldsymbol{Z})$ belongs to a known parametric family. Going further, if asymptotic validity is sufficient, we show in Section 3 that we need only the

$$\begin{aligned} &\textit{MX(2) assumption:} \text{ the first two moments of } \boldsymbol{X}|\boldsymbol{Z} \text{ are known, i.e.} \\ &\mathbb{E}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z}) \text{ and } \text{Var}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \Sigma(\boldsymbol{Z}) \text{ for known } \mu(\cdot), \Sigma(\cdot). \end{aligned} \tag{7}$$

We show that the CRT, paired with the *generalized covariance measure* statistic of Shah and Peters [1], retains asymptotic Type-I error control under the MX(2) assumption. Requiring knowledge of just the first two moments of the conditional distribution $\boldsymbol{X}|\boldsymbol{Z}$, rather than the distribution itself, promises to broaden the scope of application of MX-style methodology.

A3: The prediction error of the ML method impacts the asymptotic efficiency of the CRT but not its consistency. It has been widely observed that the better the ML method approximates $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$, the higher power the MX method will have. We put this empirical knowledge on a theoretical foundation by expressing the asymptotic power of the CRT in terms of the prediction error of the underlying ML method (Section 4). In particular, we consider semiparametric alternatives of the form

$$H_1 : \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}^T\beta + g(\boldsymbol{Z}), \sigma^2). \tag{8}$$

We analyze the power of a CRT variant that employs a separately trained estimate $\widehat{g}$ in an asymptotic regime where $d = \dim(\boldsymbol{X})$ remains fixed while $p = \dim(\boldsymbol{Z})$ grows arbitrarily with the sample size n. We find that this test is consistent no matter what $\widehat{g}$ is used, while its asymptotic power against local alternatives $\beta_n = h/\sqrt{n}$ depends on the limiting mean-squared prediction error of $\widehat{g}$ (denoted $\mathcal{E}^2$) and the limiting expected variance $\mathbb{E}[\mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}]]$ (denoted s^2). For example, if $d = 1$,

CRT power converges to that of normal location test under alternative $N\left(\frac{hs}{\sqrt{\sigma^2 + \mathcal{E}^2}}, 1\right)$.

This represents the first explicit quantification of the impact of ML prediction error on the power of an MX method.

These advances shed new light on the nature of the MX problem and can inform methodological design. Our results handle multivariate $\boldsymbol{X}$, arbitrarily correlated designs in the model for $\boldsymbol{X}$, and any black-box machine learning method to learn $\widehat{g}$.

Notation. Recalling equations (1) and (2), population-level variables (such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$) are denoted in boldface, while samples of these variables (such as X_i, Y_i, Z_i) are denoted in regular font. All vectors are treated as column vectors. We often use uppercase symbols to denote both random variables and their realizations (for either population- or sample-level quantities), but use lowercase to denote the latter when it is important to make this distinction. We use $\mathcal{L}$ to denote the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, though we sometimes use this symbol to denote the joint distribution of (X, Y, Z) as well. We use the symbol "$\equiv$" for definitions. We denote by $c_{d,1-\alpha}$ the $1-\alpha$ quantile of the χ^2_d distribution, and by $\chi^2_d(\lambda)$ the non-central χ^2 distribution with d degrees of freedom and noncentrality parameter λ.

2 The most powerful test against point alternatives

In this section, we seek the most powerful MX CI test against a point alternative. To accomplish this, we observe that any (marginally) level α test must also have level α conditionally on Y, Z (Section 2.1). The latter conditioning step reduces the composite null to a point null. This reduction allows us to invoke the Neyman Pearson lemma to find the most powerful test (Section 2.2). Proofs are deferred to Appendix A.

2.1 Conditioning reduces the composite null to a point null

Let us first formalize the definition of a level α test of the MX CI problem. The null hypothesis is defined as the set of joint distributions compatible with conditional independence and with the assumed model for $\boldsymbol{X}|\boldsymbol{Z}$:

$$\begin{aligned}\mathscr{L}_0^{\mathrm{MX}}(f^*) &\equiv \mathscr{L}_0 \cap \mathscr{L}^{\mathrm{MX}}(f^*) \\ &\equiv \{\mathcal{L} : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}\} \cap \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}}\} \\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = f_{\boldsymbol{Z}} \cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}} \cdot f_{\boldsymbol{Y}|\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}, f_{\boldsymbol{Y}|\boldsymbol{Z}}\}.\end{aligned} \tag{9}$$

A test $\phi : (\mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p)^n \to [0, 1]$ of the MX CI problem is level α if

$$\sup_{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z)] \leq \alpha. \tag{10}$$

Since the CRT calibrates the test statistic T conditionally on the observed Y, Z (recall definition (4)), it is easy to verify that this test not only has level α in the sense of equation (10) but also has level α *conditionally* on Y and Z:

$$\sup_{y,z} \sup_{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi_T^{\mathrm{CRT}}(X, Y, Z)|Y = y, Z = z] \leq \alpha. \tag{11}$$

It turns out that any level α test ϕ has this same property:[2]

Proposition 1. *Any level α test ϕ of conditional independence under the MX assumption (i.e. any test satisfying property* (10)*) must also have conditional level α:*

$$\sup_{y,z} \sup_{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z)|Y = y, Z = z] \leq \alpha. \tag{12}$$

Proposition 1 allows us to reframe the MX CI problem as that of testing a null hypothesis with respect to the conditional law $\mathcal{L}(X|Y, Z)$, viewing only X as random while fixing Y and Z at their observed values. It states that any level α test ϕ, when viewed as a *family* of hypothesis tests $\phi(X; y, z)$ indexed by (y, z), has level α in the conditional

[2]We thank Michael Celentano for pointing out this fact to us.

testing problem for each (y, z). This conditional perspective is useful because it reduces the composite null (1) to a point null. Indeed, note that

$$\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*) \Longrightarrow \mathcal{L}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i). \tag{13}$$

Therefore, $\mathcal{L}(X|Y = y, Z = z)$ equals a fixed product distribution for any null $\mathcal{L}$. This yields a conditional point null hypothesis. Note that the observations X_i are independent *but not identically distributed* due to the different conditioning events in (13).

The preceding observations will not be surprising to anyone familiar with MX methodology, and in fact the existence of a single null distribution from which to resample $\widetilde{X}$ is central to the very definition of the CRT. Nevertheless, we find it useful to state explicitly what has thus far been largely left implicit. This conditional perspective allows us to derive the optimal test against a point alternative by applying the Neyman-Pearson lemma in the conditional problem.

2.2 Most powerful test against point alternatives

The following theorem states that the CRT based on the likelihood with respect to the (unknown) distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ is the most powerful test against a point alternative. To prepare for the statement, fix an alternative distribution $\bar{\mathcal{L}} \in \mathscr{L}^{\mathrm{MX}}(f^*)$, and let $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ be the density of $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$.

Theorem 1. *Let $\bar{\mathcal{L}} \in \mathscr{L}^{MX}(f^*)$ be an alternative distribution, with $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. The likelihood of the data (X, Y, Z) with respect to $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ is*

$$T^{\mathrm{opt}}(X, Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{14}$$

The CRT $\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}$ based on this test statistic is the most powerful test of $H_0 : \mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)$ against $H_1 : \mathcal{L} = \bar{\mathcal{L}}$, i.e.

$$\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X, Y, Z)] \leq \mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}(X, Y, Z)] \tag{15}$$

for any level α test ϕ.

The proof of Theorem 1 is based on the reduction in Section 2.1 of the composite null to a point null by conditioning. This argument has similar flavor to the theory of unbiased testing (see Lehmann and Romano [16, Chapter 4]), where uniformly most powerful unbiased tests can be found by conditioning on sufficient statistics for nuisance parameters. Our result is also analogous to but different from Lehmann's derivation of the most powerful permutation tests using conditioning followed by the Neyman-Pearson lemma, in randomization-based causal inference (see the rejoinder of Rosenbaum's 2002 discussion paper [17], Section 5.10 of Lehmann (1986), now Lehmann and Romano [16, Section 5.9]).

Inspecting the most powerful test given by Theorem 1, we find that it depends on $\bar{\mathcal{L}}$ only through $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$. This immediately yields the following corollary.

Corollary 1. *Define the composite class of alternatives*

$$\begin{aligned}\mathscr{L}_1(f^*, \bar{f}) &= \{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*) : \bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}\} \\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = f_{\boldsymbol{Z}} \cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}} \cdot \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}\}.\end{aligned}$$

The CRT $\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}$ *is the uniformly most powerful test of* $H_0 : \mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)$ *against* $H_1 : \mathcal{L} \in \mathscr{L}_1(f^*, \bar{f})$.

Theorem 1 and Corollary 1 state that the most powerful test against a point alternative is the CRT based on the test statistic defined as the measuring how well the data (X, Y, Z) fit the distribution $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$. For example, if

$$\bar{f}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}^T\beta + \boldsymbol{Z}^T\gamma, \sigma^2) \text{ for coefficients } \beta \in \mathbb{R}^d \text{ and } \gamma \in \mathbb{R}^p, \tag{16}$$

then the optimal test rejects for small values of $\|Y - X\beta - Z\gamma\|^2$. In Section 5, we establish an analogous optimality statement for MX knockoffs as well. Since the optimal test depends on the alternative distribution $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, CRT and MX knockoffs implementations usually employ a machine learning step to search through the composite alternative (not unlike a likelihood ratio test) for a good approximation $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. These approximate models are then summarized in various ways to define a test statistic T. There is no consensus yet on the best test statistic to use, with some authors [2, 3, 9] using combinations of fitted coefficients $\widehat{\beta}$ and others [10, 11] using loss-based test statistics. The above optimality results align more closely with the latter strategy. Loss-based test statistics also have the advantage of avoiding ad hoc combination rules for $\widehat{\beta} \in \mathbb{R}^d$ in cases where $d > 1$. It remains to be seen whether loss-based or coefficient-based test statistics yield greater power in practice.

Intuitively, the results of this section suggest that the more successful $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ is at approximating the true alternative $f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$, the more powerful the corresponding CRT will be. We make this relationship precise in an asymptotic setting in Section 4. Before examining the asymptotic power of the CRT, we establish in the next section that asymptotic Type-I error control can be obtained under a weaker form of the MX assumption.

3 Weakening the MX assumption

Instead of assuming knowledge of the entire conditional distribution $\boldsymbol{X}|\boldsymbol{Z}$, assume only

$$\text{the conditional mean } \mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] \text{ and variance } \mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}] \text{ are known.} \tag{17}$$

We call this the *MX(2) assumption*. In this section, we show that asymptotic Type-I error can be uniformly controlled under this drastically weaker assumption (together with moment assumptions). We work in an asymptotic regime described by Setting 1 below.

Setting 1 (Arbitrary dimension asymptotics). For each $n = 1, 2, \ldots$, we have a joint law $\mathcal{L}_n$ over $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{d+r+p}$, where $d = \dim(\boldsymbol{X})$ remains fixed, $r = \dim(\boldsymbol{Y}) = 1$, and $p = \dim(\boldsymbol{Z})$ can vary arbitrarily with n. Under the MX(2) assumption, the law $\mathcal{L}_n$ has known mean and variance functions

$$\mu_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \text{ and } \Sigma_n(\boldsymbol{Z}) \equiv \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]. \tag{18}$$

For each n, we receive n i.i.d. samples $(X, Y, Z) = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ from $\mathcal{L}_n$. Note that we leave implicit the dependence on n of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ and (X, Y, Z) to lighten the notation.

By analogy with definition (9), the MX(2) null hypothesis is defined as

$$\mathscr{L}_0^{\mathrm{MX}(2)} = \mathscr{L}_0^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \equiv \mathscr{L}_0 \cap \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)), \tag{19}$$

where

$$\mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \equiv \{\mathcal{L}_n : \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = \mu_n(\boldsymbol{Z}),\ \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = \Sigma_n(\boldsymbol{Z})\}.$$

In Section 3.1, we propose a test of this MX(2) null hypothesis called the *MX(2) F-test*. We establish in Section 3.2 that this test controls Type-I error uniformly over subsets of $\mathscr{L}_0^{\mathrm{MX}(2)}$ satisfying moment conditions and that this test is asymptotically equivalent to the CRT based on the same test statistic (proofs deferred to Appendix B). We compare these results to existing ones in Section 3.3, and then check their finite-sample accuracy via numerical simulations in Section 3.4.

3.1 The MX(2) F-test

Suppose we have trained an estimate $\widehat{g}_n$ of $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$ on an independent dataset (whose size can vary arbitrarily with n and is not included in the sample size n used for testing). In the next section, g_n will denote be the nonparametric portion of a semiparametric model (33). *These training sets across n and resulting estimates $\widehat{g}_n$ remain fixed throughout.* Importantly, the sample used for training $\widehat{g}_n$ need not come from the same distribution as $\mathcal{L}_n$ and can therefore be much larger. For example, in the context of the digital twin test [11] it was pointed out that large observational datasets can be used to train $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$, while applying a variant of the CRT to a smaller experimental dataset to obtain a causal guarantee.

With the estimate $\widehat{g}_n$ in hand, it is natural to base inference on the sample covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after adjusting for $\boldsymbol{Z}$:

$$\widehat{\rho}_n \equiv \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)). \tag{20}$$

In general, $\widehat{\rho}_n \in \mathbb{R}^d$, but for $d = 1$, this coincides with the *generalized covariance measure*, proposed by Shah and Peters [1] for conditional independence testing. Related quantities

also have been studied in the semiparametric [18, 19] and doubly robust [20, 21, 22] estimation contexts; see Section 3.3 for more discussion. Constructing an asymptotically valid CI test based on $\widehat{\rho}_n$ requires us to be able to consistently estimate the limiting mean and variance of this quantity under the null. If we have access to the first two moments of $\boldsymbol{X}|\boldsymbol{Z}$, we can compute for any $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)}$ that

$$\begin{aligned}
\mathrm{Var}_{\mathcal{L}_n}[\sqrt{n}\widehat{\rho}_n] &= \mathrm{Var}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))] \\
&= \mathrm{Var}_{\mathcal{L}_n}\left[\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))|\boldsymbol{Y}, \boldsymbol{Z}]\right] + \\
&\quad \mathbb{E}_{\mathcal{L}_n}\left[\mathrm{Var}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))|\boldsymbol{Y}, \boldsymbol{Z}]\right] \\
&= \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})\right] \\
&\equiv S_n^2 \in \mathbb{R}^{d \times d}.
\end{aligned} \tag{21}$$

A natural estimate of this limiting variance is

$$\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i). \tag{22}$$

Note that we can compute both $\widehat{S}_n^2$ and $\widehat{\rho}_n$ using only the MX(2) assumption.

Based on the variance calculation (21), we would expect the standardized quantity

$$U_n(X, Y, Z) \equiv \widehat{S}_n^{-1}\sqrt{n}\widehat{\rho}_n = \frac{\widehat{S}_n^{-1}}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \in \mathbb{R}^d. \tag{23}$$

to converge to $N(0, I_d)$ under the MX(2) null (19). This motivates the *MX(2) F-test*

$$\phi_n^{\mathrm{MX}(2)}(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > c_{d,1-\alpha}),$$

where

$$T_n(X, Y, Z) \equiv \|U_n(X, Y, Z)\|^2. \tag{24}$$

See also Algorithm 1, and recall that $c_{d,1-\alpha}$ is defined as the $1-\alpha$ quantile of χ_d^2.

Algorithm 1: The MX(2) *F*-test

Data: $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, $\mu_n(\cdot)$ and $\Sigma_n(\cdot)$ in (18), learning method g

1 Obtain $\widehat{g}$ by fitting g on a separate dataset;

2 Recall $\mu_n(Z_i) \equiv \mathbb{E}_{\mathcal{L}_n}[X_i|Z_i]$, set $\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))^2\Sigma_n(Z_i)$;

3 Set $U_n \equiv \frac{\widehat{S}_n^{-1}}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i))$ and $T_n = \|U_n\|^2$;

Result: MX(2) F-test asymptotic p-value $\widehat{p} \equiv \mathbb{P}[\chi_d^2 > T_n]$.

4 **Cost:** One p-dimensional model fit.

Note that a one-sided version of this test (the *MX(2) t-test*) can be defined for $d = 1$ by rejecting for large values of $U_n(X, Y, Z)$.

Next, we formally state asymptotic type-I error control for the MX(2) F-test and claim that it is asymptotically equivalent to the CRT based on the same test statistic.

3.2 Asymptotic type-I error control and equivalence to CRT

For uniform type-I error control, we must restrict the set of null distributions (19) to those satisfying the following moment conditions for fixed $c_1, c_2 > 0$:

$$\mathscr{L}_n(c_1, c_2) \equiv \{\mathcal{L}_n : \|S_n^{-1}\| \leq c_1, \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \leq c_2\}. \quad (25)$$

Theorem 2. *Under Setting 1, if $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$ for some $c_1, c_2 > 0$, then the standardized generalized covariance measure statistic $U_n(X, Y, Z)$ converges to the standard normal:*

$$U_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \quad (26)$$

Therefore, the MX(2) F-test controls Type-I error asympotically, uniformly over the above subset of $\mathscr{L}_0^{\mathrm{MX}(2)}$:

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)] \leq \alpha. \quad (27)$$

We pause to comment on Theorem 2. It implies that much less than the MX assumption is needed if one is satisfied with asymptotic Type-I error control. Obtaining the first two moments of $\boldsymbol{X}|\boldsymbol{Z}$ is of course much easier than obtaining this entire conditional distribution (unless $\boldsymbol{X}$ is binary), so the MX(2) assumption is likely to be much easier to satisfy in practice. Furthermore, the MX(2) F-test has the computational advantage of not requiring resampling to compute its critical values, which are given explicitly. In fact, *any* method not requiring the full MX assumption must bypass resampling, since just the ability to resample from $\boldsymbol{X}|\boldsymbol{Z}$ requires the MX assumption.

While the MX(2) F-test is quite different from usual MX methods on its surface, the next theorem states that it is asymptotically equivalent to the CRT based on the same test statistic.

Theorem 3. *Under Setting 1, suppose*

$$\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2) \quad (28)$$

for some $c_1, c_2 > 0$, and define $T_n(X, Y, Z)$ via equations (23) and (24). Let ϕ_n^{CRT} denote the CRT based on T_n, with threshold $C_n(Y, Z)$ defined as in equation (4). The CRT threshold converges in probability to the MX(2) threshold:

$$C_n(Y, Z) \xrightarrow{\mathcal{L}_n}_p c_{d,1-\alpha}. \quad (29)$$

Furthermore, if $T_n(X, Y, Z)$ does not accumulate near $c_{d,1-\alpha}$, i.e.

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = 0, \quad (30)$$

then the CRT is asymptotically equivalent to the MX(2) F-test:

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z) \neq \phi_n^{\mathrm{CRT}}(X, Y, Z)] = 0. \quad (31)$$

Informally, this theorem suggests that the null distribution of $U_n(X, Y, Z)$ converges to $N(0, I_d)$, even after conditioning on Y, Z. In the language of the CRT, this means that the resampling distribution of the test statistic $T_n(X, Y, Z)$ approaches χ^2_d. This is not too surprising in retrospect, since $\mathrm{Var}_{\mathcal{L}_n}[\sqrt{n}\widehat{\rho}_n | Y, Z] = \widehat{S}_n^2$. Note that this conditional normalization property holds for the specific instance of the CRT based on the statistic T_n defined in via equations (23) and (24), though other kinds of test statistics may lead to similar large-sample behavior. We can also view Theorem 3 as a robustness result. Indeed, this theorem implies that the large-sample behavior of the CRT based on T_n depends only on the first two moments of $\boldsymbol{X}|\boldsymbol{Z}$. Therefore, the CRT carried out using a misspecified distribution $\boldsymbol{X}|\boldsymbol{Z}$ that is correct up to at least the first two moments will be asymptotically equivalent to the CRT based on the correct distribution and therefore, will control Type-I error asymptotically as well.

3.3 Comparison to existing results

Theorem 2 is reminiscent of results in the literature on semi- and non-parametric inference [18, 19, 23, 22, 1]. There, the goal is to construct asymptotically valid confidence intervals and tests for functionals like $\rho_n \equiv \mathbb{E}_{\mathcal{L}_n}[\mathrm{Cov}_{\mathcal{L}_n}[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}]]$. The connection with conditional independence testing is that $\rho_n = 0$ if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$, and the converse is true under certain semi-parametric models. Inference is usually based on some version of the functional $\widehat{\rho}_n$ defined in equation (20), with $\mu_n(\boldsymbol{Z}) = \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ estimated alongside $g_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$. Asymptotically valid inference is obtained under variants of the "doubly robust" assumption that the estimates for $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$ are both consistent, with the product of the estimation errors tending to zero at a rate of $o(n^{-1/2})$. By contrast, the MX(2) F-test places more weight on the model for $\boldsymbol{X}|\boldsymbol{Z}$ (assuming both first and second moments of this conditional distribution are known) while placing less weight on the model for $\boldsymbol{Y}|\boldsymbol{Z}$ (not assuming a semi-parametric form for $\boldsymbol{Y}|\boldsymbol{Z}$ or even consistency for $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$). The estimate $\widehat{S}_n^2$ we employ for the variance of $\widehat{\rho}_n$, when compared to those typically used in the semi-parametric literature, reflects this differing set of assumptions.

Theorem 3 is a statement about the asymptotic equivalence between the resampling-based CRT and the asymptotic MX(2) F-test. The CRT is in the spirit of the finite-population approach to causal inference (Fisher), whereas the MX(2) F-test is in the spirit of the asymptotic super-population approach (Neyman). We find that research in these two strands of work on causal inference have proceeded largely separately from each other, and therefore connections between the two have received relatively little attention. However, there has been a recent line of work [24, 25, 26] focusing on the asymptotic behavior of the Fisher randomization test in the context of completely randomized experiments. A similar result to Theorem 3 is that the Fisher randomization test (analogous to the CRT) is asymptotically equivalent to the Rao score test (analogous to the MX(2) F-test) in a completely randomized experiment [24, Theorem A.1]. Theorem 3 can be viewed as an extension of this result to accommodate for non-binary treatments as well as high-dimensional covariates affecting both treatment and response.

3.4 Finite-sample convergence assessment

Theorem 2 states that $U_n(X, Y, Z)$ converges to $N(0, I_d)$ unconditionally, while Theorem 3 is related to the convergence of $U_n(X, Y, Z)$ to $N(0, I_d)$ conditionally on Y, Z. In this section, we describe a numerical simulation designed to assess both convergence statements in finite samples. Code to reproduce the simulation is available online at `https://github.com/ekatsevi/crtpower-manuscript`.

Simulation setup. Note that, if $(\boldsymbol{X}, \boldsymbol{Z})$ is jointly Gaussian, then $U_n(X, Y, Z)$ is exactly distributed as $N(0, I_d)$ both unconditionally and conditionally in finite samples; see also Section 2.5 of [12]. To test the above convergence statements in a nontrivial setting, we instead consider a discrete distribution for $(\boldsymbol{X}, \boldsymbol{Z})$. In particular, we sample $(\boldsymbol{X}, \boldsymbol{Z})$ from a Markov chain, as described next.

Let's assume for simplicity that $\dim(\boldsymbol{X}) = 1$. Define $(\boldsymbol{X}, \boldsymbol{Z}) \in \{0, 1\}^{1+p}$ to have the distribution of a Markov chain with

$$\text{initial state } \boldsymbol{X} \sim \text{Ber}(\pi_{\text{init}}) \text{ and transition matrix } \begin{pmatrix} 1 - \pi_{\text{flip}} & \pi_{\text{flip}} \\ \pi_{\text{flip}} & 1 - \pi_{\text{flip}} \end{pmatrix}.$$

More explicitly, we have

$$\mathbb{P}[\boldsymbol{X} = x, \boldsymbol{Z} = z] = \pi_{\text{init}}^{x}(1-\pi_{\text{init}})^{1-x}\pi_{\text{flip}}^{\mathbb{1}(z_1 \neq x)}(1-\pi_{\text{flip}})^{\mathbb{1}(z_1 = x)} \prod_{j=2}^{p} \pi_{\text{flip}}^{\mathbb{1}(z_j \neq z_{j-1})}(1-\pi_{\text{flip}})^{\mathbb{1}(z_j = z_{j-1})}$$

The parameters $(\pi_{\text{init}}, \pi_{\text{flip}})$ describe the distribution of $\boldsymbol{X}|\boldsymbol{Z}$ and are assumed known. Furthermore, let the response $\boldsymbol{Y}$ be distributed as a random effects model in $\boldsymbol{Z}$:

$$\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 I_p), \ \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n).$$

Thus, all simulations are conducted under the null hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. The signal-to-noise ratio in this relationship is defined via

$$\text{SNR} = \frac{\mathbb{E}[\|\boldsymbol{Z}\|^2]\sigma_\gamma^2}{\sigma_\epsilon^2}.$$

The function $\widehat{g}_n$ is defined by running a 10-fold cross-validated ridge regression of Y on Z on n_{train} training samples, and then the statistic $U_n(X, Y, Z)$ is evaluated based on n_{test} independent test samples.

Simulation parameters. All simulations were run with

$$n_{\text{train}} = 100; \quad \pi_{\text{init}} = 0.1; \quad \pi_{\text{flip}} = 0.1; \quad \sigma_\epsilon^2 = 1. \tag{32}$$

On the other hand, the three parameters $(n_{\text{test}}, \text{SNR}, p)$ were varied as follows:

$$n_{\text{test}} \in \{10, 25, \mathbf{100}\}; \quad \text{SNR} \in \{0, \mathbf{1}, 5\}; \quad p \in \{20, 100, \mathbf{500}\}.$$

The bolded values above represent the *default values* for each parameter. Each of the three parameters was varied while keeping the other two parameters at their default values, giving a total of nine simulation settings.

For each simulation setting, the training data and the estimate $\widehat{g}_n$ were generated just once, since our results condition on the training data. The entire test data (X, Y, Z) were sampled 1000 times to generate the unconditional distribution of $U_n(X, Y, Z)$. To generate the conditional distribution of this quantity, the data (Y, Z) were sampled once per problem setting, and then $X|Z$ was sampled 1000 times.

Simulation results. For each of the nine simulation settings, normal QQ plots of the unconditionally-generated $U_n(X, Y, Z)$ are shown in Figure 1 while their conditional counterparts are shown in Figure 2. Based on these results, we make the following observations. The sample size impacts calibration, but the SNR and the dimension do not. In particular, unconditional and conditional normality are already achieved at $n = 100$, regardless of SNR and dimension. Unconditional convergence to normality tends to be faster than conditional convergence, with the unconditional distribution already normal for $n = 25$ while the conditional distribution has not yet converged for this value of n. Our main conclusion is that for even moderate sample sizes, the MX(2) F-test controls Type-I error and behaves quite similarly to the CRT. We must bear in mind, however, that different choices of the fixed parameters (32) may alter these conclusions. In particular, smaller π_{init} leads to more discreteness in X and therefore slower convergence to normality.

Next, we study the asymptotic power of the MX(2) F-test (and by Theorem 3, of the CRT) against semiparametric alternatives.

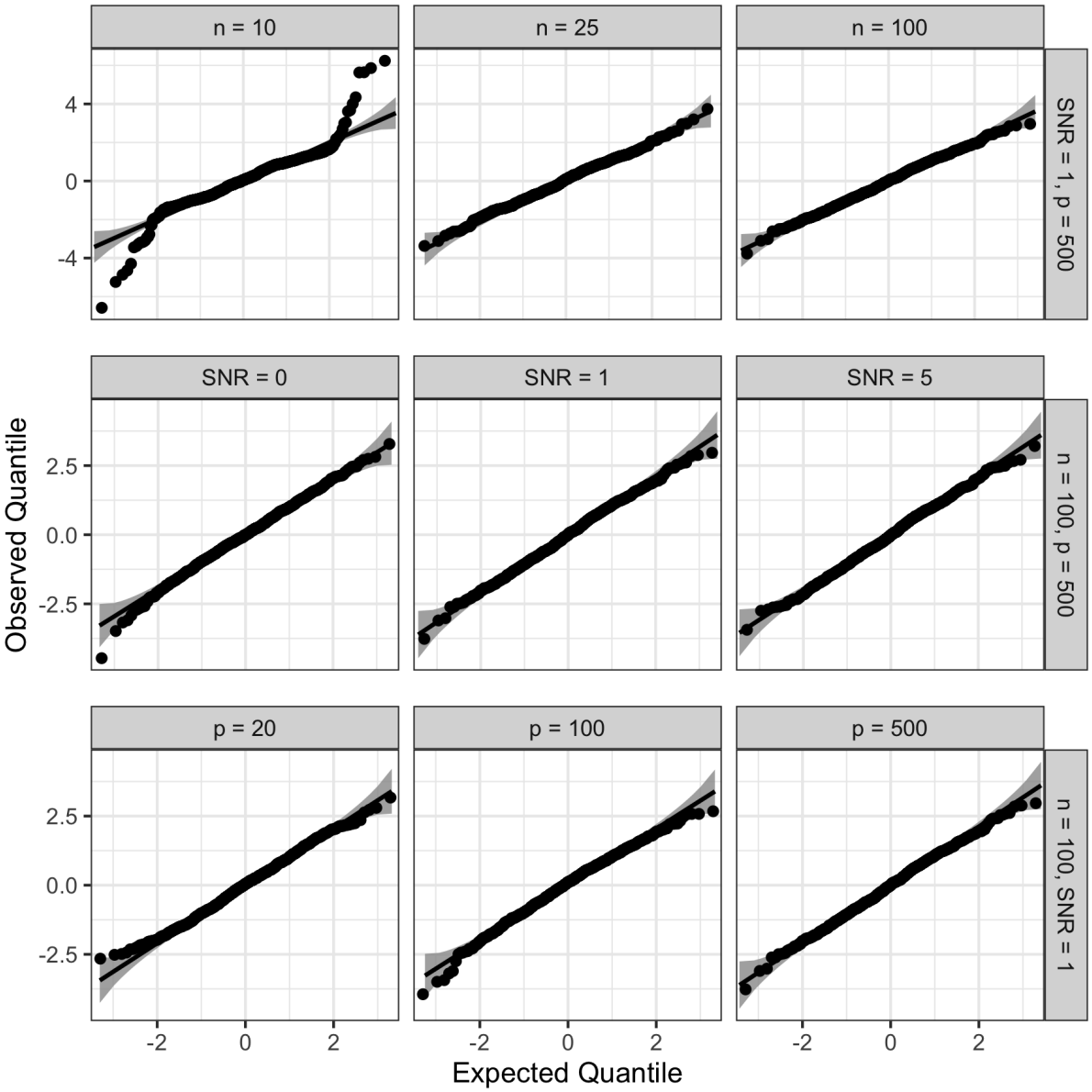

Figure 1: Unconditional distribution of $U_n(X, Y, Z)$ across the nine simulation settings.

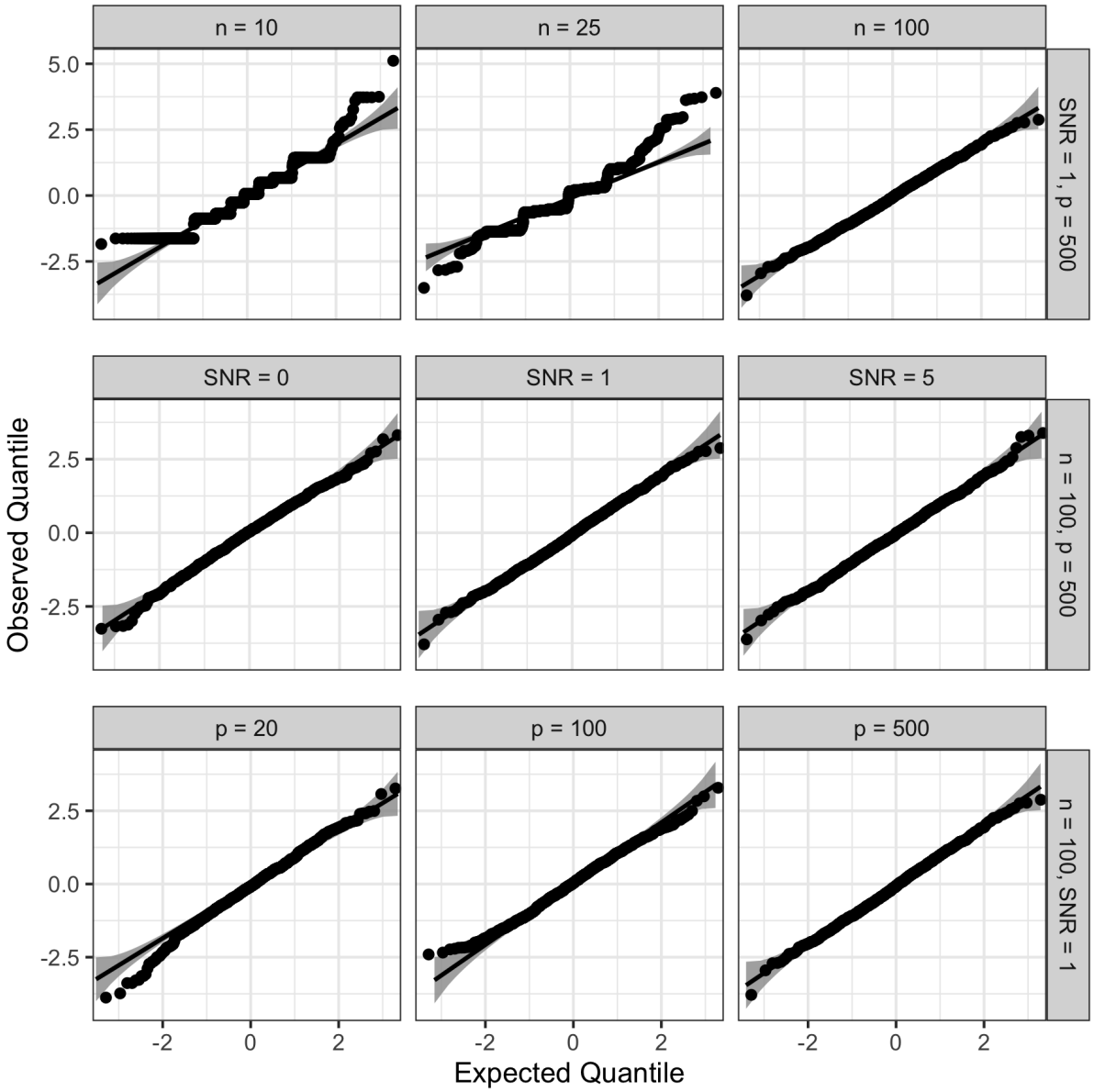

Figure 2: Conditional distribution of $U_n(X, Y, Z)|Y, Z$ across the nine simulation settings.

4 The asymptotic power of the CRT

In Section 2, we saw how to construct the optimal test against point alternatives specified by $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. In practice, of course we do not have access to this distribution, so we usually estimate it via a statistical machine learning procedure. The goal of this section is to quantitatively assess the power of the CRT as a function of the prediction error of this ML procedure.

4.1 Power against semiparametric alternatives

In this section, we consider semiparametric alternatives as described next.

Setting 2 (Semiparametric alternatives). Under Setting 1, assume $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$ is such that

$$\boldsymbol{Y} = (\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta_n + g_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(0,\sigma^2),\ \sigma^2 > 0 \tag{33}$$

for $\boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$. Here, $\beta_n \in \mathbb{R}^d$ is a coefficient vector, $g_n : \mathbb{R}^p \to \mathbb{R}$ a general function, and $\sigma^2 > 0$ the residual variance.

Note that the function $\widehat{g}_n$ from the previous section can be viewed as an approximation to $g_n(\boldsymbol{Z})$. The semiparametric model (33) has been extensively studied (see e.g. the classic works [18, 19]), but not in the context of MX methods and mostly focusing on the estimation problem.

In Theorem 4 below, we express the asymptotic power of the MX(2) F-test against alternatives (33) in terms of the variance-weighted mean square error of $\widehat{g}_n$:

$$\mathcal{E}_n^2 \equiv \mathbb{E}_{\mathcal{L}_n}\left[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^2 \cdot \overline{\Sigma}_n^{-1/2}\Sigma_n(\boldsymbol{Z})\overline{\Sigma}_n^{-1/2}\right], \quad \text{where } \overline{\Sigma}_n \equiv \mathbb{E}_{\mathcal{L}_n}[\Sigma_n(\boldsymbol{Z})]. \tag{34}$$

Note that if $(\boldsymbol{X},\boldsymbol{Z})$ is jointly Gaussian, then $\Sigma_n(\boldsymbol{Z}) = \overline{\Sigma}_n$ for all $\boldsymbol{Z}$ and therefore $\mathcal{E}_n^2 = \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^2] \cdot I_d$. Our result requires the following moment assumptions:

$$\sup_n \|\overline{\Sigma}_n^{-1}\| < \infty, \tag{35}$$

$$\sup_n\ \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^8] < \infty, \tag{36}$$

and

$$\sup_n\ \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^4\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] < \infty. \tag{37}$$

Theorem 4. *Consider semiparametric alternative Setting 2. Suppose $\mathcal{L}_n$ satisfies the moment conditions* (35)*,* (36)*, and* (37)*, and that the conditional variance and variance-weighted mean squared error converge:*

$$\overline{\Sigma}_n \to \overline{\Sigma} \quad \textit{and} \quad \mathcal{E}_n^2 \to \mathcal{E}^2 \quad \textit{as } n \to \infty, \tag{38}$$

Then, we have the following two statements:

(a) (Consistency) If $\beta_n = \beta \neq 0$ for each n, then the MX(2) F-test and the CRT based on the same statistic are consistent:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX}(2)}(X,Y,Z)\right] = \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{CRT}}(X,Y,Z)\right] = 1. \tag{39}$$

(b) (Power against local alternatives) If $\beta_n = h_n/\sqrt{n}$ for a convergent sequence $h_n \to h \in \mathbb{R}^d$, then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX}(2)}(X,Y,Z)\right] &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{CRT}}(X,Y,Z)\right] \\ &= \mathbb{P}[\chi_d^2(\|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\Sigma^{1/2}h\|^2) > c_{d,1-\alpha}].\end{aligned} \tag{40}$$

Recalling that $\chi_d^2(\lambda)$ denotes the noncentral chi-square distribution with d degrees of freedom and non-centrality parameter λ, the second part of Theorem 4 states that the MX(2) F-test and the CRT based on the same test statistic have power equal to that of a χ^2 test of a multivariate normal random vector having mean zero under the alternative $N((\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\Sigma^{1/2}h, I_d)$. This result establishes a direct link between the estimation error in $\widehat{g}_n$ and the power of the CRT against local alternatives. In particular, the mean-squared error term $\mathcal{E}^2$ contributes additively to the irreducible error term $\sigma^2 I_d$. We can gain intuition for this result by considering the regression model

$$\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}) = (\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta_n + (g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}) \tag{41}$$

obtained from the semiparametric model (33) by subtracting $\widehat{g}_n(\boldsymbol{Z})$ from both sides. The test statistic T_n is based on the quantity $\widehat{\rho}_n$ defined in equation (20), which can be viewed as an unnormalized version of the fitted regression coefficients of $Y - \widehat{g}_n(Z)$ on $X - \mu_n(Z)$. The term $g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z})$ in the regression model (41) contributes additively to the residual error term, so in a traditional regression analysis we would expect the power of the test to depend on the variance of this error term. In fact, standard large-sample OLS theory (see e.g. Section 2.3 of Hayashi's book [27]) states that the power against local alternatives of the F-test in the regression model (41) is exactly the same as that of the MX(2) F-test stated in equation (40). Of course, the usual F-test applied to the regression (41) relies on the validity of this model while the MX(2) F-test instead relies on knowledge of $\mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}]$. Note that [28] also find the power of an MX test and a classical OLS test to have the same power (see their Appendix F).

4.2 Example: Power of lasso-based CRT

A key ingredient in the power formula (40) is the limiting variance-weighted mean squared error $\mathcal{E}^2$. This error depends on the machine learning method used to obtain $\widehat{g}_n$. We can leverage existing results about the asymptotic behavior of prediction error of machine learning methods in high dimensions. In this section, we consider the case when $\widehat{g}_n$ is trained using the lasso in the orthogonal design case, which was studied by Bayati and Montanari [29].

Setting 3 (Linear regression with orthogonal design). Under Setting 2, assume further that $(\boldsymbol{X}, \boldsymbol{Z}) \sim N(0, I_{1+p})$, $\beta_n = h_n/\sqrt{n}$ for some convergent sequence $h_n \to h \in \mathbb{R}$, and $g_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\gamma_n$. Suppose $\gamma_n \in \mathbb{R}^p$ is such that the entries of $\sqrt{n}\gamma_n$ converge weakly to a random variable Γ on $\mathbb{R}$ such that $\mathbb{P}[\Gamma \neq 0] > 0$ and $\|\sqrt{n}\gamma_n\|^2/p \to \mathbb{E}[\Gamma^2] < \infty$.

Until now, we have denoted by n the sample size used for constructing tests, leaving unspecified the size of the separate sample used to train $\widehat{g}_n$. To get concrete expressions for the power of the MX(2) F-test and the CRT based on a specific machine learning method to obtain $\widehat{g}_n$, we must take the training sample size into account. We therefore define tests $\varphi_n^{\mathrm{MX(2)}}(X, Y, Z)$ and $\varphi_n^{\mathrm{CRT}}(X, Y, Z)$, which for some training proportion $\pi \in (0, 1)$ split the data into πn training observations $(X_{\mathrm{train}}, Y_{\mathrm{train}}, Z_{\mathrm{train}})$ and $(1-\pi)n$ test observations $(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}})$. These tests both proceed by first running a lasso of Y_{train} on Z_{train} with regularization parameter λ to obtain an estimate $\widehat{\gamma}_{\pi n}$. The tests $\varphi_n^{\mathrm{MX(2)}}(X, Y, Z)$ and $\varphi_n^{\mathrm{CRT}}(X, Y, Z)$ are then obtained by running the MX(2) F-test and the CRT on the test data $(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}})$ using the estimate $\widehat{g}_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\widehat{\gamma}_{\pi n}$:

$$\varphi_n^{\mathrm{MX(2)}}(X, Y, Z) \equiv \phi_{(1-\pi)n}^{\mathrm{MX(2)}}(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}}) \text{ and } \varphi_n^{\mathrm{CRT}}(X, Y, Z) \equiv \phi_{(1-\pi)n}^{\mathrm{CRT}}(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}}).$$

Note that the dependence of $\phi_{(1-\pi)n}^{\mathrm{MX(2)}}(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}})$ and $\phi_{(1-\pi)n}^{\mathrm{CRT}}(X_{\mathrm{test}}, Y_{\mathrm{test}}, Z_{\mathrm{test}})$ on the training data $(X_{\mathrm{train}}, Y_{\mathrm{train}}, Z_{\mathrm{train}})$ is left implicit.

Under Setting 3, we can directly use Bayati and Montanari's theory [29] to obtain

$$\lim_{n\to\infty} \mathcal{E}_n^2 = \tau_*^2 - \sigma^2 \quad \text{a.s. in } (X_{\mathrm{train}}, Y_{\mathrm{train}}, Z_{\mathrm{train}}), \tag{42}$$

where (α_*, τ_*) is the unique solution of the system below:

$$\begin{aligned} \lambda &= \alpha\tau(1 - (\pi\delta)^{-1}\mathbb{E}[\eta'(\sqrt{\pi}\Gamma + \tau W; \alpha\tau)]) \\ \tau^2 &= \sigma^2 + (\pi\delta)^{-1}\mathbb{E}[(\eta(\sqrt{\pi}\Gamma + \tau W; \alpha\tau) - \sqrt{\pi}\Gamma)^2]. \end{aligned} \tag{43}$$

Here, $W \sim N(0, 1)$ is independent of Γ and $\eta(x; \theta) = (|x| - \theta)_+\mathrm{sign}(x)$ is the soft threshold function. This leads to the following corollary of Theorem 4, proved in Appendix B:

Corollary 2. *Under Setting 3, the asymptotic power of φ_n^{CRT} and $\varphi_n^{\mathrm{MX(2)}}$ converges to that of a standard normal location test with alternative mean $\tau_*^{-1}h\sqrt{1-\pi}$:*

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\mathrm{CRT}}(X, Y, Z)] &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\mathrm{MX(2)}}(X, Y, Z)] \\ &= \mathbb{P}[|N(\tau_*^{-1}h\sqrt{1-\pi}, 1)| > z_{1-\alpha/2}]. \end{aligned} \tag{44}$$

Corollary 2 gives the power of these lasso-based methods in a very simple form, with the prediction error of the lasso entering through the effective noise level τ_*. The impact of the splitting proportion π on power can be seen in the multiplication of the signal strength h by $\sqrt{1-\pi}$. The splitting proportion implicitly impacts the effective noise level τ_* as well; smaller π lead to greater effective noise levels. Note that the expectations in Corollary 2 are over both training and test sets, while the expectations in Theorem 4 are over the test set only.

4.3 Comparison to existing results

Two other power analyses of the CRT have been recently conducted [28, 30], focusing on the case where $g_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\gamma_n$, $\widehat{g}_n$ is trained using the lasso, $n/p \to \delta$, and the generalized covariance measure test statistic $\widehat{\rho}_n$ is used. The former study considers the case of orthogonal design (Setting 3), while the latter considers arbitrary joint Gaussian distribution for $(\boldsymbol{X}, \boldsymbol{Z})$. Assuming $\mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]] \to s^2$ (the quantity we called $\overline{\Sigma}$ in Section 4.1, with different notation to clarify that for $\dim(\boldsymbol{X}) = 1$ the covariance matrix simply becomes a variance), the works [28, 30] found that the power of the CRT with in-sample lasso fit tends to that of a normal location test with alternative mean sh/τ_*, where τ_* is the effective noise level from AMP theory (in the orthogonal design case, (α_*, τ_*) are defined by equation (43) with $\pi = 1$).

This is a similar expression to what we found in Corollary 2 in the orthogonal design case. Furthermore, note that $\tau_*^2 = \sigma^2 + \mathcal{E}^2$ (i.e. the out-of-sample prediction error of the lasso). It follows that the power expression found by [28, 30] is exactly the same as what we found in part (b) of Theorem 4, despite the fact that $\widehat{g}_n$ is fit in-sample. The power of CRT with $\widehat{g}_n$ fit in-sample is a more subtle object to analyze, as it may depend on the degree to which $\widehat{g}_n$ overfits. [28] also derive a power expression for the CRT when $\widehat{g}_n$ is fit in-sample via ordinary least squares (allowing correlated covariates, as we do in Setting 2), which also happens to coincide with the expression (40). Such in-sample results have only been obtained only for these two test statistics, however, though it would be interesting whether expression (40) holds for more general classes of test statistics. By contrast, training $\widehat{g}_n$ on a separate sample allows us to prove Theorem 4 for very broad classes of machine learning methods $\widehat{g}_n$.

Finally, we note a connection between Theorem 4 and causal inference. It is widely known in causal inference (see e.g. [31, Section 7.5]) that adjustment for covariates Z in randomized experiments (a) yields consistent estimates despite misspecification of $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and (b) improve estimation efficiency to the extent that this adjustment captures the distribution $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$. This fact mirrors the conclusions of Theorem 4. The asymptotic variance of the regression-based estimator for the average treatment effect in a completely randomized experiment is a standard result, but we are unaware of a quantitative expression of the asympotic efficiency of covariate-adjusted versions of the Fisher randomization test (though some insight is provided by [26]).

5 The most powerful one-bit p-values for knockoffs

MX knockoffs [2] operate differently than the CRT; they simultaneously test the conditional associations of many variables with a response. Given m variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and a response $\boldsymbol{Y}$, it is of interest to test the CI hypotheses

$$H_j : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \quad j = 1, \ldots, m.$$

Note that j indexes variables, rather than samples. Comparing to our setup, $\boldsymbol{X}_j$ plays the role of $\boldsymbol{X}$ and $\boldsymbol{X}_{-j}$ plays the role of $\boldsymbol{Z}$. In particular, we allow $\boldsymbol{X}_j$ to be a group of

variables. Like HRT, knockoffs only requires one model fit, so it too is computationally faster than the CRT. Among these three MX procedures, knockoffs is currently the most popular. We briefly review it next, and then present an optimality result in the spirit of Theorem 1. Its proof is given in the supplement (Section C).

5.1 A brief overview of knockoffs

A set of knockoff variables $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_m)$ is constructed to satisfy conditional exchangeability:

$$\mathcal{L}(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}) = \mathcal{L}(\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}), \quad j = 1, \ldots, m \tag{45}$$

and conditional independence

$$\boldsymbol{Y} \perp\!\!\!\perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{X}. \tag{46}$$

Given such a construction, a set of knockoff variables $\widetilde{X}_{i,\bullet}$ is sampled from $\mathcal{L}(\widetilde{\boldsymbol{X}} | \boldsymbol{X} = X_{i,\bullet})$ for each i. Knockoff inference is then based on a form of data-carving: variables are given an ordering $\tau(1), \ldots, \tau(m)$ determined arbitrarily from $([X, \widetilde{X}], Y)$ as long as $X_{\bullet,j}$ and $\widetilde{X}_{\bullet,j}$ are treated symmetrically, and then tested in that order based on *one-bit p-values* p_j measuring the contrast between the strength of association between $X_{\bullet,j}$ and Y and that between $\widetilde{X}_{\bullet,j}$ and Y. Given any statistic $T_j([X, \widetilde{X}], Y)$ measuring the strength of association between X_j and Y, define the one-bit p-value

$$p_j([X, \widetilde{X}], Y) \equiv \begin{cases} \frac{1}{2}, & \text{if } T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y); \\ 1, & \text{if } T_j([X, \widetilde{X}], Y) \leq T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y). \end{cases} \tag{47}$$

Here, $[X, \widetilde{X}]_{\text{swap}(j)}$ is defined as the result of swapping $X_{\bullet,j}$ with $\widetilde{X}_{\bullet,j}$ in $[X, \widetilde{X}]$ while keeping all other columns in place. A set of variables with guaranteed false discovery rate control is chosen via the ordered testing procedure *Selective SeqStep*, applied to the p-values p_j in the order τ.

5.2 The most powerful one-bit p-value

It is harder to analyze the power of knockoffs than that of the CRT for several reasons. Knockoffs is fundamentally a *multiple* testing procedure, coupling the analysis of H_j across variables j. Furthermore, the qualities of the ordering τ and of the one-bit p-values p_j both contribute to the power of knockoffs. Due to these challenges, no optimality results are currently available for knockoffs. We take a first step in this direction by exhibiting the test statistics T_j that lead to the most powerful one-bit p-values against a point alternative.

Theorem 5. *Let $\bar{\mathcal{L}}$ be a fixed alternative distribution for $(\boldsymbol{X}, \boldsymbol{Y})$, with $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}) = \bar{f}(\boldsymbol{Y}|\boldsymbol{X})$. Define the likelihood statistic*

$$T_j^{\text{opt}}([X, \widetilde{X}], Y) \equiv \prod_{i=1}^{n} \bar{f}(Y_i | X_{i,\bullet}). \tag{48}$$

Assuming that ties do not occur, that is

$$\mathbb{P}_{\bar{\mathcal{L}}}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] = 0, \tag{49}$$

we have that the above likelihood statistic yields the optimal one-bit p-value:

$$T_j^{\text{opt}} \in \arg\max_{T_j} \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y)]. \tag{50}$$

The reader observes that the optimal test statistic is not a function of the knockoff variables, which may seem paradoxical. Recall from the definition (47), however, that the one-bit p-value compares the test statistic on the original and swapped augmented design $[X, \widetilde{X}]$. Therefore, the optimal one-bit p-value checks whether the original jth variable $X_{\bullet,j}$ fits with the rest of the data better than does its knockoff $\widetilde{X}_{\bullet,j}$. A simple way of operationalizing Theorem 5 is to fit a model $\widehat{f}(\boldsymbol{Y}|\boldsymbol{X})$ based on $([X, \widetilde{X}], Y)$ in any way that treats original variables and knockoffs symmetrically, and then defining $T_j([X, \widetilde{X}], Y) \equiv \widehat{f}(Y|X)$. The above result continues to hold when $\boldsymbol{X}_j$ is a *group* of variables, giving a clean way to combine evidence across multiple variables. A conditional version of the optimality statement (50) holds; see equation (104) in the supplement.

Theorem 5 requires that ties occur with probability zero (49). Proposition 2 below states that this nondegeneracy condition holds if either $\boldsymbol{Y}|\boldsymbol{X}$ or $\boldsymbol{X}_j|\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has continuous distribution.

Proposition 2. *Suppose $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}) = g_{\boldsymbol{\eta}}$, where $\boldsymbol{\eta} = \boldsymbol{X}_j\beta_j + f_{-j}(\boldsymbol{X}_{-j})$ and g_η is a one-dimensional exponential family with natural parameter η and strictly convex, continuous log partition function ψ. Suppose also that $\boldsymbol{X}_j, \beta_j \in \mathbb{R}$, with $\beta_j \neq 0$. The nondegeneracy condition* (49) *holds if either*

1. *$\boldsymbol{X}_j|\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has a density for each $\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$, or*
2. *g_η has a density,*

where the densities are with respect to the Lebesgue measure.

Finally, we remark that there are a few existing power analyses for knockoffs, all in high-dimensional asymptotic regimes and assuming lasso-based test statistics. Weinstein et al [32] analyze the power of a knockoffs variant in the case of independent Gaussian covariates, while Liu and Rigollet [33] and Fan et al [34] study conditions for consistency under correlated designs. Our finite-sample optimality result is complimentary to these previous works.

6 Discussion

In this paper, we gave some answers to the theoretical questions posed in the introduction. We presented the first finite-sample optimality results in the MX framework, exhibited a

significantly weakened form of the MX assumption and a methodology valid under only this assumption, and explicitly quantified how the performance of the underlying ML procedure impacts the asymptotic power of the CRT.

The MX framework is just one setting where black-box prediction methods have been recently employed for the purpose of more powerful statistical inference. Other examples include conformal prediction [35], classification-based two-sample testing [36] and data-carving based multiple testing [37]. These methods employ ML algorithms to create powerful test statistics, calibrating them for valid inference with no assumptions about the method used. However, the more accurate the learned model, the more powerful the inference. Our finite-sample and asymptotic power results explicitly tie the error of the learning algorithm to the power of the test, and thus put this common intuition on a quantitative foundation and may thus help inform the choice and design of ML methods used for inferential goals.

Another set of connections we highlighted throughout the paper is to causal inference and semiparametric estimation. The MX CI problem has strong similarities to the problem of testing Fisher's strong null in a randomized experiment with potentially non-binary treatment and known propensity function. Furthermore, the CRT is similar in spirit to the Fisher randomization test. We believe these connections can be further leveraged to address problems in the MX framework that remain open. For example, consider the situation when the MX assumption is only approximately correct. This is analogous to the situation in observational studies, where the propensity score/function must be estimated. There is a vast literature on this topic based on "double robustness/machine-learning" [22] or targeted learning [23]. Similar ideas may help relax the MX assumption [8] or study robustness to its misspecification [38]. Another topic that has received little attention in the MX community is that of estimation (with the exception of [39]). Causal inference is a rich source of meaningful estimands (such as the *dose response function* [40]) and estimators (such as the proposal of Kennedy et al. [41] for doubly-robust dose response function estimation). Such ideas may be directly relevant to the MX framework.

Much still remains to be done to systematically understand the theoretical properties of MX methods. One interesting direction is to analyze the case when $\widehat{g}_n$ is learned on the same data as is used for testing. We saw in Section 4.3 that Theorem 4 extends to lasso-based estimators $\widehat{g}_n$ learned in-sample, but the generality of such results remains an open question. It would also be interesting to consider alternatives beyond the linear model (33). A natural next step would be to consider generalized linear models. Furthermore, the connections to causal inference referenced above are tantalizing and deserve a dedicated treatment. Finally, we hope that these new theoretical insights about MX methods will lead to improved methodologies that are both statistically and computationally efficient, along the lines of the CRT variants discussed in this paper and in recent work [12].

Acknowledgments

We thank Asaf Weinstein, Timothy Barry, and Stephen Bates for detailed comments on earlier versions of the manuscript, as well as Ed Kennedy and Larry Wasserman for discussions of the connections to causal inference.

References

[1] Rajen D. Shah and Jonas Peters. "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure". In: *Annals of Statistics, to appear* (2020). arXiv: 1804.07203.

[2] Emmanuel Candès et al. "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 551–577.

[3] M. Sesia, C. Sabatti, and E. J. Candès. "Gene hunting with hidden Markov model knockoffs". In: *Biometrika* 106.1 (2019), pp. 1–18.

[4] Kosuke Imai and David A. Van Dyk. "Causal inference with general treatment regimes: Generalizing the propensity score". In: *Journal of the American Statistical Association* 99.467 (2004), pp. 854–866.

[5] Paul R. Rosenbaum and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[6] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. "Deep Knockoffs". In: *Journal of the American Statistical Association* 0.0 (2019), pp. 1–27.

[7] Stephen Bates et al. "Metropolized Knockoff Sampling". In: *Journal of the American Statistical Association* (2020). arXiv: 1903.00434v1.

[8] Dongming Huang and Lucas Janson. "Relaxing the Assumptions of Knockoffs by Conditioning". In: *Annals of Statistics, to appear* (2020). arXiv: 1903.02806.

[9] Matteo Sesia et al. "Multi-resolution localization of causal variants across the genome". In: *Nature Communications* 11 (2020), p. 1093.

[10] Wesley Tansey et al. "The Holdout Randomization Test: Principled and Easy Black Box Feature Selection". In: *arXiv* (2018). arXiv: 1811.00645.

[11] Stephen Bates et al. "Causal Inference in Genetic Trio Studies". In: *arXiv* (2020). arXiv: arXiv:2002.09644v1.

[12] Molei Liu et al. "Fast and Powerful Conditional Randomization Testing via Distillation". In: *arXiv* (2020).

[13] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1 (1996), pp. 267–288.

[14] Lu Zheng and Marvin Zelen. "Multi-center clinical trials: Randomization and ancillary statistics". In: *Annals of Applied Statistics* 2.2 (2008), pp. 582–600.

[15] Jonathan Hennessy et al. "A Conditional Randomization Test to Account for Covariate Imbalance in Randomized Experiments". In: *Journal of Causal Inference* 4.1 (2016), pp. 61–80.

[16] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Third. New York: Springer, 2005.

[17] Paul R. Rosenbaum. "Covariance adjustment in randomized experiments and observational studies". In: *Statistical Science* 17.3 (2002), pp. 286–327.

[18] P. M. Robinson. "Root-N-Consistent Semiparametric Regression". In: *Econometrica* 56.4 (1988), pp. 931–954.

[19] James M. Robins, Steven D. Mark, and Whitney K. Newey. "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders". In: *Biometrics* 48.2 (1992), pp. 479–495.

[20] James M. Robins and Andrea Rotnitzky. "Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer"". In: *Statistica Sinica* 11.4 (2001), pp. 920–936.

[21] Mark J. van der Laan and James M. Robins. *Unified methods for censored longitudinal data and causality*. New York: Springer-Verlag, 2003.

[22] Victor Chernozhukov et al. "Double/debiased machine learning for treatment and structural parameters". In: *Econometrics Journal* 21.1 (2018), pp. C1–C68.

[23] Mark J. van der Laan and Sherri Rose. *Targeted learning: Causal inference for observational and experimental data*. New York: Springer, 2011.

[24] Peng Ding. "A paradox from randomization-based causal inference". In: *Statistical Science* 32.3 (2017), pp. 331–345. arXiv: 1402.0142.

[25] Jason Wu and Peng Ding. "Randomization Tests for Weak Null Hypotheses in Randomized Experiments". In: *Journal of the American Statistical Association* (2020). arXiv: 1809.07419.

[26] Anqi Zhao and Peng Ding. "Covariate-adjusted Fisher randomization tests for the average treatment effect". In: *Journal of Econometrics* 94720 (2021). arXiv: 2010.14555.

[27] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

[28] Wenshuo Wang and Lucas Janson. "A Power Analysis of the Conditional Randomization Test and Knockoffs." In: *arXiv* (2020).

[29] Mohsen Bayati and Andrea Montanari. "The LASSO risk for Gaussian matrices". In: *IEEE Transactions on Information Theory* 58.4 (2011), pp. 1997–2017. arXiv: 1008.2581.

[30] Michael Celentano, Andrea Montanari, and Yuting Wei. "The Lasso with general Gaussian designs with applications to hypothesis testing". In: *arXiv* (2020). arXiv: arXiv:2007.13716v1.

[31] Guido W. Imbens and Donald B. Rubin. *Causal inference: For statistics, social, and biomedical sciences an introduction*. Cambridge University Press, 2015.

[32] Asaf Weinstein, Rina Barber, and Emmanuel Candes. "A power analysis for knockoffs under Gaussian designs". In: *arXiv* (2017). arXiv: 1712.06465.

[33] Jingbo Liu and Philippe Rigollet. "Power analysis of knockoff filters for correlated designs". In: *33rd Conference on Neural Information Processing Systems*. 2019. arXiv: 1910.12428.

[34] Yingying Fan et al. "RANK: Large-Scale Inference With Graphical Nonlinear Knockoffs". In: *Journal of the American Statistical Association* 115.529 (2020), pp. 362–379. arXiv: 1709.00092.

[35] Rina Foygel Barber et al. "Predictive inference with the jackknife+". In: *Annals of Statistics, to appear* (2020).

[36] Ilmun Kim et al. "Classification accuracy as a proxy for two sample testing". In: *Annals of Statistics, to appear* (2020). arXiv: 1602.02210.

[37] Lihua Lei and William Fithian. "AdaPT: an interactive procedure for multiple testing with side information". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 649–679.

[38] Rina Foygel Barber, Emmanuel J. Candès, and Richard J. Samworth. "Robust inference with knockoffs". In: *Annals of Statistics, to appear* (2020). arXiv: 1801.03896.

[39] Lu Zhang and Lucas Janson. "Floodgate : inference for model-free variable importance". In: *arXiv* (2020), pp. 1–67.

[40] Keisuke Hirano and Guido W. Imbens. "The Propensity Score with Continuous Treatments". In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (2004), pp. 73–84.

[41] Edward H Kennedy et al. "Non-parametric methods for doubly robust estimation of continuous treatment effects". In: *Journal of the Royal Statistical Society, Series B (Methodological)* 4 (2017), pp. 1229–1245.

[42] Robert Lang. "A note on the measurability of convex sets". In: *Archiv der Mathematik* 47.1 (1986), pp. 90–92.

A Proofs for Section 2

Proof of Proposition 1. Fix (y, z) and $\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)$. Defining

$$\mathcal{L}_{y,z}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \equiv \delta_{(\boldsymbol{Y},\boldsymbol{Z})=(y,z)} \cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}} \in \mathscr{L}_0^{\mathrm{MX}}(f^*),$$

note that

$$\mathbb{E}_{\mathcal{L}}[\phi(X,Y,Z)|Y=y,Z=z] = \int \phi(x,y,z) f^*_{\boldsymbol{X}|\boldsymbol{Z}}(x|z)dx = \mathbb{E}_{\mathcal{L}_{y,z}}[\phi(X,Y,Z)] \leq \alpha.$$

This completes the proof. □

Proof of Theorem 1. Fix realizations y, z. We first claim that $\phi^{\text{CRT}}_{T^{\text{opt}}}$ is the most powerful test in the conditional problem, i.e.

$$\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,y,z)|Y=y,Z=z] \leq \mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\text{CRT}}_{T^{\text{opt}}}(X,y,z)|Y=y,Z=z] \tag{51}$$

for any test $\phi(X,y,z)$ satisfying

$$\sup_{\mathcal{L}\in\mathscr{L}_0^{\text{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X,y,z)|Y=y,Z=z] \leq \alpha. \tag{52}$$

In the conditional problem, the alternative $\bar{\mathcal{L}}$ induces the following distribution for X:

$$\bar{\mathcal{L}}(X=x|Y=y,Z=z) = \prod_{i=1}^n f^*(x_i|z_i)\tfrac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)} \tag{53}$$

where

$$\bar{f}(y_i|z_i) \equiv \int \bar{f}(y_i|x_i,z_i) f^*(x_i|z_i)dx_i.$$

The conditional problem is therefore a test of

$$H_0 : \mathcal{L}(X=x|Y=y,Z=z) = \prod_{i=1}^n f^*(x_i|z_i) \quad \text{versus}$$
$$H_1 : \mathcal{L}(X=x|Y=y,Z=z) = \prod_{i=1}^n f^*(x_i|z_i)\tfrac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)}.$$

This is a simple testing problem, with point null and point alternative. By the Neyman-Pearson lemma, the most powerful test is the one that rejects for large values of the likelihood ratio

$$\prod_{i=1}^n \frac{P_1(x_i|y_i,z_i)}{P_0(x_i|y_i,z_i)} = \prod_{i=1}^n \frac{f^*(x_i|z_i)\frac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)}}{f^*(x_i|z_i)} = \prod_{i=1}^n \frac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)} \propto T^{\text{opt}}(x,y,z),$$

verifying the conditional optimality claim (51). To obtain the unconditional optimality claim (15), note that by Proposition 1 any unconditionally level α test ϕ must also have level α in the conditional problem (52). For any such test, we may therefore conclude

$$\begin{aligned}\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,Y,Z)] &= \mathbb{E}_{\bar{\mathcal{L}}}[\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,Y,Z)|Y,Z]] \\ &\leq \mathbb{E}_{\bar{\mathcal{L}}}\left[\mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\text{CRT}}_{T^{\text{opt}}}(X,Y,Z)|Y,Z]\right] = \mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\text{CRT}}_{T^{\text{opt}}}(X,Y,Z)],\end{aligned}$$

as desired. □

B Proofs for Sections 3 and 4

B.1 Proofs of main results

Proof of Theorem 2. Fix any sequence $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$. Because $\mathcal{L}_n \in \mathscr{L}_0$, we have $(X, Y, Z) \overset{d}{=} (\widetilde{X}, Y, Z)$, where $\widetilde{X}_i | Y, Z \overset{\text{ind}}{\sim} \mathcal{L}_n(\boldsymbol{X} | \boldsymbol{Z} = Z_i)$. By conclusion (87) of Lemma 3, which applies because $\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$ by assumption, we have $U_n(X, Y, Z) \overset{d}{=} U_n(\widetilde{X}, Y, Z) \overset{\mathcal{L}_n}{\to}_d N(0, I_d)$. This verifies the asymptotic normality statement (26).

To show the asymptotic Type-I error control statement (27), it suffices to show that for any sequence $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$, we have

$$\limsup_{n \to \infty} \ \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)] \leq \alpha. \tag{54}$$

By the continuous mapping theorem it follows from asymptotic normality (26) that $T_n(X, Y, Z) = \|U_n(X, Y, Z)\|^2 \overset{\mathcal{L}_n}{\to}_d \chi_d^2$. Therefore,

$$\lim_{n \to \infty} \ \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)] = \lim_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[T_n(X, Y, Z) > c_{d,1-\alpha}] = \mathbb{P}[\chi_d^2 > c_{d,1-\alpha}] = \alpha,$$

from which the conclusion (54) follows. This completes the proof. □

Proof of Theorem 3. First, conclusion (87) of Lemma 3—which applies because of the assumption $\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$—states that for

$$\widetilde{X}_i^1, \widetilde{X}_i^2 | Y, Z \overset{\text{ind}}{\sim} \mathcal{L}_n(\boldsymbol{X} | \boldsymbol{Z} = Z_i),$$

we have the convergence

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z) \end{pmatrix} \overset{\mathcal{L}_n}{\to}_d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix} \right). \tag{55}$$

By the continuous mapping theorem, we find that

$$(T_n(\widetilde{X}^1, Y, Z), T_n(\widetilde{X}^2, Y, Z)) \overset{\mathcal{L}_n}{\to}_d \chi_d^2 \times \chi_d^2. \tag{56}$$

Since χ_d^2 has a continuous and strictly increasing distribution function, we conclude using Lemma 1 that $C_n(Y, Z) \overset{\mathcal{L}_n}{\to}_p Q_{1-\alpha}[\chi_d^2] = c_{d,1-\alpha}$, proving the statement (29).

Next, note that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z) \neq \phi_n^{\mathrm{CRT}}(X, Y, Z)] \\
&\quad = \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z))] \\
&\quad = \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z)), |C_n(Y, Z) - c_{d,1-\alpha}| \leq \delta] \\
&\qquad + \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z)), |C_n(Y, Z) - c_{d,1-\alpha}| > \delta] \\
&\quad \leq \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] + \mathbb{P}_{\mathcal{L}_n}[|C_n(Y, Z) - c_{d,1-\alpha}| > \delta].
\end{aligned}$$

To justify the last step, suppose without loss of generality that $c_{d,1-\alpha} \leq C_n(Y,Z)$. Then, note that if $c_{d,1-\alpha} < T_n(X,Y,Z) \leq C_n(Y,Z)$ and $C_n(Y,Z) - c_{d,1-\alpha} \leq \delta$ then

$$|T_n(X,Y,Z) - c_{d,1-\alpha}| = T_n(X,Y,Z) - c_{d,1-\alpha} \leq C_n(Y,Z) - c_{d,1-\alpha} \leq \delta.$$

Taking a lim sup on both sides in the display before the last and using the convergence $C_n(Y,Z) \xrightarrow{\mathcal{L}_n}_p c_{d,1-\alpha}$, we find that

$$\begin{aligned}
&\limsup_{n\to\infty} \ \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X,Y,Z) \neq \phi_n^{\mathrm{CRT}}(X,Y,Z)] \\
&\quad \leq \limsup_{n\to\infty} \ \mathbb{P}_{\mathcal{L}_n}[|T_n(X,Y,Z) - c_{d,1-\alpha}| \leq \delta] + \limsup_{n\to\infty} \ \mathbb{P}_{\mathcal{L}_n}[|C_n(Y,Z) - c_{d,1-\alpha}| > \delta] \\
&\quad = \limsup_{n\to\infty} \ \mathbb{P}_{\mathcal{L}_n}[|T_n(X,Y,Z) - c_{d,1-\alpha}| \leq \delta].
\end{aligned}$$

Letting $\delta \to 0$ and using the assumption (30), we arrive at the claimed asymptotic equivalence (31). This completes the proof. □

Proof of Theorem 4. We start by proving consistency. To this end, we claim that

$$\widehat{\rho}_n \xrightarrow{\mathcal{L}_n}_p \overline{\Sigma}\beta. \tag{57}$$

Indeed, $\widehat{\rho}_n$ is the mean of i.i.d. terms with expectation

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta + \boldsymbol{\epsilon} + g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))] = \overline{\Sigma}_n\beta.
\end{aligned} \tag{58}$$

These terms also have bounded second moment, since

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))\|^2] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta + \boldsymbol{\epsilon} + g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\
&\quad \leq C\mathbb{E}_{\mathcal{L}_n}[(\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2\|\beta\|^2 + \boldsymbol{\epsilon}^2 + (g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2)\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\
&\quad = C\|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] + C\sigma^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\
&\qquad + C\mathbb{E}_{\mathcal{L}_n}[(g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2].
\end{aligned} \tag{59}$$

Here, C is a constant so that $(a+b+c)^2 \leq C(a^2+b^2+c^2)$ for any $a, b, c \geq 0$. Taking a supremum over n and using the assumptions (36) and (37) yields

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))\|^2] < \infty. \tag{60}$$

Therefore, the weak law of large numbers implies that

$$\widehat{\rho}_n - \overline{\Sigma}_n\beta \xrightarrow{\mathcal{L}_n}_p 0, \tag{61}$$

from which the statement (57) follows by the assumed convergence $\overline{\Sigma}_n \to \overline{\Sigma}$. Next, we derive that

$$T_n(X,Y,Z) = \|\sqrt{n}\widehat{S}_n^{-1}\widehat{\rho}_n\|^2 = \|\sqrt{n}\widehat{S}_n^{-1}S_nS_n^{-1}\widehat{\rho}_n\|^2 \geq \left(\sqrt{n}\lambda_{\min}(\widehat{S}_n^{-1}S_n)\lambda_{\min}(S_n^{-1})\|\widehat{\rho}_n\|\right)^2.$$

Now, we have $\widehat{S}_n^{-1} S_n \xrightarrow{\mathcal{L}_n}_p I_d$ by conclusion (80) of Lemma 2, so the continuous mapping theorem implies that $\lambda_{\min}(\widehat{S}_n^{-1} S_n) \xrightarrow{\mathcal{L}_n}_p 1$. Furthermore, $\inf_n \lambda_{\min}(S_n^{-1}) > 0$ by conclusion (94) of Lemma 4. Finally, $\|\widehat{\rho}_n\| \xrightarrow{\mathcal{L}_n}_p \|\overline{\Sigma}\beta\|$ by equation (57), and

$$\|\overline{\Sigma}\beta\| \geq \lambda_{\min}(\overline{\Sigma})\|\beta\| = \|\overline{\Sigma}^{-1}\|^{-1}\|\beta\| \geq \left(\sup_n \|\overline{\Sigma}_n^{-1}\|\right)^{-1} \|\beta\| > 0$$

since $\beta \neq 0$ by assumption and assumptions (35) and (38) imply that $\|\overline{\Sigma}^{-1}\| \leq \sup_n \|\overline{\Sigma}_n^{-1}\| < \infty$. Putting these facts together implies that $\sqrt{n}\lambda_{\min}(\widehat{S}_n^{-1} S_n)\lambda_{\min}(S_n^{-1})\|\widehat{\rho}_n\| \xrightarrow{\mathcal{L}_n}_p \infty$, and therefore $T_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_p \infty$. Hence,

$$\mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)] = \mathbb{P}_{\mathcal{L}_n}[T_n(X, Y, Z) > c_{d,1-\alpha}] \to 1. \tag{62}$$

The fact that $T_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_p \infty$ also implies that $\limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = 0$ for any $\delta > 0$. Hence, the condition (30) of Theorem 3 is satisfied, so the conclusion (31) implies that

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{CRT}}(X, Y, Z)] = \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)] = 1.$$

Thus, we have shown the claimed consistency (39), so we have finished the proof of part (a) of the theorem.

To prove part (b), we claim that it suffices to establish that

$$T_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d \chi_d^2(\|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2} h\|^2). \tag{63}$$

Indeed, the limiting power of the MX(2) F-test would directly follow from this statement. To establish that the CRT has the same limiting power, by Theorem 3 it suffices to verify the non-accumulation condition (30). Letting T be the limiting distribution in claim (63), this claim implies that for any $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = \mathbb{P}[|T - c_{d,1-\alpha}| \leq \delta].$$

Because T has a continuous distribution function, the limit above tends to zero. Therefore, it is indeed sufficient to verify the claimed convergence (63). This statement, in turn, will follow if we prove that

$$U_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d N((\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma}h, I_d). \tag{64}$$

Indeed, note that

$$h^T\overline{\Sigma}(\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1}\overline{\Sigma}h = h^T\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)^{-1}\overline{\Sigma}^{1/2}h = \|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2}h\|^2.$$

To show the statement (64), we first rewrite $U_n(X, Y, Z)$ as follows:

$$\begin{aligned}
&U_n(X, Y, Z) \\
&= \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^{n} ((X_i - \mu_n(Z_i))^T \beta_n + \epsilon_i + g_n(Z_i) - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \\
&= \frac{\widehat{S}_n^{-1}}{n} \sum_{i=1}^{n} (X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T h_n + \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \\
&\equiv A_n + B_n,
\end{aligned}$$

where $Y_i' \equiv g_n(Z_i) + \epsilon_i$. It therefore suffices to show that

$$A_n \xrightarrow{\mathcal{L}_n}_p (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma}h \quad \text{and} \quad B_n \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{65}$$

By conclusion (93) of Lemma 4, there exist c_1, c_2 for which $\mathcal{L}_n \in \mathscr{L}(c_1, c_2)$ for each n. Therefore, we can apply Lemma 2 to conclude that

$$\widehat{S}_n^{-1} S_n \xrightarrow{\mathcal{L}_n}_p I_d. \tag{66}$$

By conclusion (95) of Lemma 4, we have that $S_n^2 \to \overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2}$, so

$$S_n^{-1} \to (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}. \tag{67}$$

Now, we apply the WLLN to find the limit of A_n. Since $(X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T$ has expectation $\overline{\Sigma}$ and second moment uniformly bounded by the eighth moment assumption (36), we can apply the weak law of large numbers as well as the statements (66) and (67) to conclude that

$$A_n = (\widehat{S}_n^{-1} S_n)\frac{S_n^{-1}}{n} \sum_{i=1}^{n} (X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T h_n \xrightarrow{\mathcal{L}_n}_p (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma}h.$$

Next, we seek to find the limit of B_n. Defining $\boldsymbol{Y}', S_n'^2, \mathcal{L}_n'$ according to (92) below, we may rewrite

$$B_n = (\widehat{S}_n^{-1} S_n)(S_n^{-1} S_n')\frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)). \tag{68}$$

By conclusion (95) of Lemma 4, we have

$$S_n^{-1} S_n' \to I_d. \tag{69}$$

Furthermore, conclusion (93) of Lemma 4 gives $\mathcal{L}_n' \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$. Therefore, $\mathcal{L}_n'$ satisfies the assumptions of Lemma 3, statement (86) of which gives

$$\frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(\widetilde{X}_i - \mu_n(Z_i)) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{70}$$

Furthermore, $\mathcal{L}'_n \in \mathscr{L}_0$ implies that $(\boldsymbol{X}, \boldsymbol{Y}', \boldsymbol{Z}) \stackrel{d}{=} (\widetilde{\boldsymbol{X}}, \boldsymbol{Y}', \boldsymbol{Z})$, which together with the convergence (70) implies that

$$\frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \stackrel{\mathcal{L}_\eta}{\to}_d N(0, I_d). \tag{71}$$

Finally, putting together displays (66), (69) and (71) yields that $B_n \stackrel{\mathcal{L}_\eta}{\to}_d N(0, I_d)$. This verifies the claimed convergences (65) and therefore completes the proof. □

Proof of Corollary 2. First we verify the statement (42). To this end, first note that

$$\begin{aligned}
\mathcal{E}_n^2 &= \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^2 \overline{\Sigma}_n^{-1/2} \Sigma_n(\boldsymbol{Z}) \overline{\Sigma}_n^{-1/2} \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^2 \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\widehat{\gamma}_{\pi n} - \gamma_n)^T \boldsymbol{Z}\boldsymbol{Z}^T (\widehat{\gamma}_{\pi n} - \gamma_n) \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= (\widehat{\gamma}_{\pi n} - \gamma_n)^T \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Z}\boldsymbol{Z}^T] (\widehat{\gamma}_{\pi n} - \gamma_n) \\
&= \|\widehat{\gamma}_{\pi n} - \gamma_n\|^2.
\end{aligned} \tag{72}$$

The second equality holds because for $(\boldsymbol{X}, \boldsymbol{Z})$ jointly Gaussian, $\Sigma_n(\boldsymbol{Z})$ is constant in $\boldsymbol{Z}$, so $\overline{\Sigma}_n^{-1/2} \Sigma_n(\boldsymbol{Z}) \overline{\Sigma}_n^{-1/2} = 1$. Therefore, the variance-weighted mean-squared error $\mathcal{E}_n^2$ of $\widehat{g}_n$ reduces to the squared error in the estimate $\widehat{\gamma}_{\pi n}$. To obtain the limit of the latter quantity, we appeal to Bayati and Montanari's Corollary 1.6 [29]. To verify the conditions of this corollary, it suffices to verify part (b) of their Definition 1: that the empirical distribution of the noise terms $\epsilon_i' \equiv Y_i - Z_i^T \gamma_n = X_i \beta_n + \epsilon_i$ in the training set (say $1 \le i \le \pi n$) converges weakly to a random variable Λ and $\frac{1}{\pi n} \sum_{i=1}^{\pi n} \epsilon_i'^2 \to \mathbb{E}[\Lambda^2]$. These statements hold almost surely in the training data by the strong law of large numbers if we assume without loss of generality that X_i and ϵ_i are both defined as the first πn elements of infinite i.i.d. sequences with distributions $N(0,1)$ and $N(0, \sigma^2)$, respectively. Therefore, Bayati and Montanari's Corollary 1.6 gives

$$\frac{1}{p} \|\sqrt{\pi n} \widehat{\gamma}_{\pi n} - \sqrt{\pi n} \gamma_n\|^2 \stackrel{\text{a.s.}}{\to} \pi\delta(\tau_*^2 - \sigma^2), \tag{73}$$

where the almost sure statement is with respect to the training data $(X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}})$. Since $\pi n / p \to \pi\delta$, we can cancel these terms from the above equation to obtain

$$\|\widehat{\gamma}_{\pi n} - \gamma_n\|^2 \stackrel{\text{a.s.}}{\to} \tau_*^2 - \sigma^2. \tag{74}$$

Putting together equations (72) and (74) gives the claimed statement (42).

To apply this result, we must verify the assumptions of Theorem 4. The bounded inverse assumption (35) holds because $\overline{\Sigma}_n = 1$ for all n in the orthogonal design setting. The eighth moment assumption (36) holds due to the boundedness of the eighth moments of Gaussian random variables. To verify the moment assumption (37), we note that,

almost surely in the training data,

$$\begin{aligned}
&\sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^4 \|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&\quad = \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Z}\widehat{\gamma}_{\pi n} - \boldsymbol{Z}\gamma_n)^4 \|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&\quad \leq \sup_n \|\widehat{\gamma}_{\pi n} - \gamma_n\|^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{Z}\|^4 \|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] \\
&\quad \leq \sup_n \|\widehat{\gamma}_{\pi n} - \gamma_n\|^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{Z}\|^8]^{1/2} \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^8]^{1/2} \\
&\quad < \infty.
\end{aligned}$$

The last inequality holds because $\|\widehat{\gamma}_{\pi n} - \gamma_n\|^4$ has a finite limit according to (74) and because $\boldsymbol{Z}$ and $\boldsymbol{X}$ have bounded eighth moments since they are Gaussian. Finally, we verify assumption (38) by noting that $\overline{\Sigma}_n \to \overline{\Sigma} \equiv 1$ and $\mathcal{E}_n^2 \to \tau_*^2 - \sigma^2 \equiv \mathcal{E}$, the latter by statement (42). Therefore, Theorem 4 gives

$$\mathbb{E}_{\mathcal{L}_n}[\phi^{\text{MX}(2)}_{(1-\pi)n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}}) | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \overset{\text{a.s.}}{\to} \mathbb{P}[\chi_1^2(\|\tau_*^{-1} h\sqrt{1-\pi}\|^2) > c_{1,1-\alpha}]$$

and

$$\mathbb{E}_{\mathcal{L}_n}[\phi^{\text{CRT}}_{(1-\pi)n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}}) | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \overset{\text{a.s.}}{\to} \mathbb{P}[\chi_1^2(\|\tau_*^{-1} h\sqrt{1-\pi}\|^2) > c_{1,1-\alpha}].$$

The extra factor of $\sqrt{1-\pi}$ reflects the fact that a sample size of $(1-\pi)n$ is used for testing, so $\beta_n = h_n/\sqrt{n} = h_n\sqrt{1-\pi}/\sqrt{(1-\pi)n}$. In other words, reducing the number of samples for testing from n to $(1-\pi)n$ has the effect of reducing the alternative signal strength from h_n to $h_n\sqrt{1-\pi}$. Noting that $c_{1,1-\alpha} = z_{1-\alpha/2}^2$, we conclude using the dominated convergence theorem that

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\text{MX}(2)}(X, Y, Z)] &= \mathbb{E}_{\mathcal{L}_n}\left[\mathbb{E}_{\mathcal{L}_n}[\phi^{\text{MX}(2)}_{(1-\pi)n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}}) | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}]\right] \\
&\to \mathbb{P}[|N(\tau_*^{-1} h\sqrt{1-\pi}, 1)| > z_{1-\alpha/2}],
\end{aligned}$$

and likewise that

$$\mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\text{CRT}}(X, Y, Z)] \to \mathbb{P}[|N(\tau_*^{-1} h\sqrt{1-\pi}, 1)| > z_{1-\alpha/2}].$$

This completes the proof of the corollary. □

B.2 Technical lemmas

First, we state a lemma that gives a sufficient condition for the convergence of the CRT threshold, which follows directly from Lemmas 2 and 3 of [28].

Lemma 1 ([28])**.** *Let $\mathcal{L}_n$ be a sequence of laws over $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, from which (X, Y, Z) are sampled. Furthermore, for each i sample two independent copies*

$$\widetilde{X}_i^1, \widetilde{X}_i^2 \overset{i.i.d.}{\sim} \mathcal{L}_n(\boldsymbol{X} | \boldsymbol{Z} = Z_i) \quad \textit{such that, given } Z, (\widetilde{X}_1^1, \widetilde{X}_1^2) \perp\!\!\!\perp \cdots \perp\!\!\!\perp (\widetilde{X}_n^1, \widetilde{X}_n^2) \perp\!\!\!\perp Y. \quad (75)$$

Suppose that $T_n(X, Y, Z)$ is a test statistic satisfying

$$(T_n(\widetilde{X}^1, Y, Z), T_n(\widetilde{X}^2, Y, Z)) \xrightarrow{\mathcal{L}_n}_d \widetilde{T} \times \widetilde{T} \tag{76}$$

for some limiting random variable $\widetilde{T}$ with continuous and strictly increasing distribution function. Then, the CRT threshold converges in probability to the upper quantile of $\widetilde{T}$:

$$C_n(Y, Z) \equiv Q_{1-\alpha}[T_n(\widetilde{X}, Y, Z)|Y, Z] \xrightarrow{\mathcal{L}_n}_p Q_{1-\alpha}[\widetilde{T}]. \tag{77}$$

Lemma 2. *Fix any $c_1, c_2 > 0$. For any sequence*

$$\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2), \tag{78}$$

we have

$$\widehat{S}_n^2 - S_n^2 \xrightarrow{\mathcal{L}_n}_p 0 \tag{79}$$

and

$$\widehat{S}_n^{-1} S_n \xrightarrow{\mathcal{L}_n}_p I_d. \tag{80}$$

Proof. To show the first convergence (79), we apply the WLLN to the triangular array $\{(Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i)\}_{i,n}$. We first verify the second moment condition:

$$\begin{aligned}
&\sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})\|^2] \\
&\quad = \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \|\Sigma_n(\boldsymbol{Z})\|^2] \\
&\quad \leq \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2 | \boldsymbol{Z}]^2] \\
&\quad \leq \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]] \\
&\quad \leq c_2 < \infty.
\end{aligned} \tag{81}$$

Therefore, by the WLLN we obtain the convergence

$$\widehat{S}_n^2 - S_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i) - \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \xrightarrow{\mathcal{L}_n}_p 0. \tag{82}$$

To show the second convergence (80), note first that

$$\begin{aligned}
\sup_n \ \|S_n^2\| &= \sup_n \ \|\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})]\| \\
&\leq \sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})\|^2]^{1/2} \leq c_2^{1/2},
\end{aligned} \tag{83}$$

the last step having been derived in equation (81). Therefore, for every n, we have

$$S_n^2 \in \mathcal{S} \equiv \{S^2 : \|S^{-1}\| \leq c_1, \|S^2\| \leq c_2^{1/2}\}. \tag{84}$$

Since $\mathcal{S}$ is a compact subset of the open set of positive definite matrices, there exists a $\delta > 0$ such that $\mathcal{S}_\delta = \{S^2 : \|S^2 - S_0^2\| \leq \delta \text{ for some } S_0^2 \in \mathcal{S}\}$ is also a compact subset of the set of positive definite matrices. Since the function $S^2 \mapsto S^{-1}$ is continuous on the compact set $\mathcal{S}_\delta$, it must be uniformly continuous on this set as well. Fix $\gamma > 0$. By uniform continuity, there exists an $\eta > 0$ such that $\|S_1^2 - S_2^2\| \leq \eta$ implies that $\|S_1^{-1} - S_2^{-1}\| \leq \gamma$ for all $S_1^2, S_2^2 \in \mathcal{S}_\delta$. We therefore have that

$$\begin{aligned}\mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma] &= \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma, \widehat{S}_n^2 \in \mathcal{S}_\delta] + \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma, \widehat{S}_n^2 \notin \mathcal{S}_\delta] \\ &\leq \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \eta] + \mathbb{P}_{\mathcal{L}_n}[\widehat{S}_n^2 \notin \mathcal{S}_\delta] \\ &\leq \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \eta] + \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \delta].\end{aligned}$$

Using the convergence (79), we find that the last expression tends to zero as $n \to \infty$, from which it follows that $\mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma] \to 0$ as $n \to \infty$. Therefore,

$$\widehat{S}_n^{-1} - S_n^{-1} \overset{\mathcal{L}_n}{\to}_p 0.$$

Multiplying this relation on the right by the bounded quantity S_n, we arrive at the statement (80), which concludes the proof. □

Lemma 3. *Consider generating* $(\widetilde{X}^1, \widetilde{X}^2, Y, Z)$ *according to* (75) *for a sequence of laws*

$$\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2). \tag{85}$$

We have

$$n^{-1/2} \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i)) \begin{pmatrix} \widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i) \end{pmatrix} \overset{\mathcal{L}_n}{\to}_d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix} \right) \tag{86}$$

and

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z) \end{pmatrix} \overset{\mathcal{L}_n}{\to}_d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix} \right). \tag{87}$$

Proof of Lemma 3. Note that

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z) \end{pmatrix} = \begin{pmatrix} \widehat{S}_n^{-1} S_n & 0 \\ 0 & \widehat{S}_n^{-1} S_n \end{pmatrix} \cdot n^{-1/2} \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i)) \begin{pmatrix} \widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i) \end{pmatrix}.$$

By Lemma 2, we have that $\widehat{S}_n^{-1} S_n \overset{\mathcal{L}_n}{\to}_p I_d$, so by Slutsky we find that the second statement (87) follows from the first (86). Therefore, it suffices to prove the latter convergence. To this end, we apply the CLT to the triangular array of vectors

$$\left\{ (Y_i - \widehat{g}_n(Z_i)) \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i) \end{pmatrix} \right\}_{i,n}. \tag{88}$$

To apply the CLT, we first verify the Lyapunov condition with $\delta = 1$:

$$\begin{aligned}
&\sup_n \ \mathbb{E}_{\mathcal{L}_n}\left[\left\|(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))\begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1}\end{pmatrix}\begin{pmatrix}\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2-\mu_n(\boldsymbol{Z})\end{pmatrix}\right\|^3\right] \\
&\le \sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[|\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z})|^3(\|\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z})\|^2+\|\widetilde{\boldsymbol{X}}^2-\mu_n(\boldsymbol{Z})\|^2)^{3/2}\right] \\
&\le \sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[|\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z})|^3 C\left(\|\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z})\|^3+\|\widetilde{\boldsymbol{X}}^2-\mu_n(\boldsymbol{Z})\|^3\right)\right] \\
&= 2C\sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[|\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z})|^3\|\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z})\|^3\right] \\
&\le 2C\sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\|\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z})\|^4\right]^{3/4} \\
&= 2C\sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\mathbb{E}_{\mathcal{L}_n}[\|\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]\right]^{3/4} \\
&= 2C\sup_n \ \|S_n^{-1}\|^3\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]\right]^{3/4} \\
&\le 2Cc_1^3c_2^{3/4} < \infty.
\end{aligned} \tag{89}$$

Here C is chosen such that $(a+b)^{3/2} \le C(a^{3/2}+b^{3/2})$ for all $a, b \ge 0$. Next, it is easy to verify that

$$\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))\begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1}\end{pmatrix}\begin{pmatrix}\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2-\mu_n(\boldsymbol{Z})\end{pmatrix}\right] = \begin{pmatrix}0\\0\end{pmatrix} \tag{90}$$

and

$$\mathrm{Var}_{\mathcal{L}_n}\left[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))\begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1}\end{pmatrix}\begin{pmatrix}\widetilde{\boldsymbol{X}}^1-\mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2-\mu_n(\boldsymbol{Z})\end{pmatrix}\right] = \begin{pmatrix} I_d & 0 \\ 0 & I_d\end{pmatrix}. \tag{91}$$

By the CLT, the convergence (86) now follows. □

Lemma 4. *In the setting of Theorem 4, define*

$$\boldsymbol{Y}' \equiv g_n(\boldsymbol{Z})+\boldsymbol{\epsilon}, \quad S_n'^2 \equiv \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}'-\widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})], \quad \textit{and} \quad \mathcal{L}_n' \equiv \mathcal{L}_n(\boldsymbol{X},\boldsymbol{Y}',\boldsymbol{Z}). \tag{92}$$

Under the assumptions of Theorem 4(a) or 4(b),

$$\textit{there exist } c_1, c_2 > 0 \textit{ such that } \mathcal{L}_n, \mathcal{L}_n' \in \mathscr{L}(c_1, c_2). \tag{93}$$

Under the assumptions of Theorem 4(a), we have

$$\inf_n \ \lambda_{\min}(S_n^{-1}) > 0, \tag{94}$$

while under the assumptions of Theorem 4(b), we have

$$\lim_{n\to\infty} S_n^2 = \lim_{n\to\infty} S_n'^2 = \overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2}. \tag{95}$$

Proof. First, we show that under the assumptions of Theorem 4(a) or 4(b), we have $\mathcal{L}_n \in \mathscr{L}(c_1, c_2)$ for some $c_1, c_2 > 0$. It suffices to show that

$$\sup_n \|S_n^{-1}\| < \infty \tag{96}$$

and

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] < \infty. \tag{97}$$

To show the statement (96), first note that

$$\begin{aligned}
S_n^2 &= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n + \boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^2 \Sigma_n(\boldsymbol{Z})] \\
&\quad + 2\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n (\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad + \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^2 \Sigma_n(\boldsymbol{Z})] + \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})],
\end{aligned} \tag{98}$$

where in the last step we used the fact that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n (\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))|\boldsymbol{Z}]^T \beta_n (g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] = 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
S_n'^2 &= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{\epsilon} + g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{\epsilon}^2 \Sigma_n(\boldsymbol{Z})] + \mathbb{E}_{\mathcal{L}_n}[2\boldsymbol{\epsilon}(g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad + \mathbb{E}_{\mathcal{L}_n}[(g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \sigma^2 \overline{\Sigma}_n + \overline{\Sigma}_n^{1/2} \mathcal{E}_n^2 \overline{\Sigma}_n^{1/2} = \overline{\Sigma}_n^{1/2} (\sigma^2 I_d + \mathcal{E}_n^2) \overline{\Sigma}_n^{1/2}.
\end{aligned} \tag{99}$$

It follows that $S_n^2 \succcurlyeq \sigma^2 \overline{\Sigma}_n$, which together with assumption (35) implies that

$$\sup_n \|S_n^{-1}\| \leq \sup_n \|\sigma^{-1} \overline{\Sigma}_n^{-1/2}\| = \sigma^{-1} \left(\sup_n \|\overline{\Sigma}_n^{-1}\|\right)^{1/2} < \infty. \tag{100}$$

This verifies statement (96). To prove statement (97), we write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}\left[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n + \boldsymbol{\epsilon} + g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad \leq C\mathbb{E}_{\mathcal{L}_n}\left[(((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^4 + \boldsymbol{\epsilon}^4 + (g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4) \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad \leq C\|\beta_n\|^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^8] + 3C\sigma^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] \\
&\qquad + C\mathbb{E}_{\mathcal{L}_n}\left[(g_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4 \|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4\right].
\end{aligned}$$

Here, C a constant such that $(a+b+c)^4 \leq C(a^4+b^4+c^4)$ for all $a, b, c \geq 0$. Taking a supremum over n and using the moment assumptions (36) and (37) along with the boundedness of the sequence β_n yields the statement (97).

Therefore, we have verified that $\mathcal{L}_n \in \mathscr{L}(c_1, c_2)$ for some c_1, c_2 under the assumptions of Theorem 4(a) or 4(b). The fact that $\mathcal{L}_n' \in \mathscr{L}(c_1, c_2)$ under these assumptions follows by a similar argument (omitted for the sake of brevity), which finishes the proof of statement (93).

Next, we turn to proving the claim (94). Using calculations (98) and (99), we write

$$S_n^2 = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta)^2\Sigma_n(\boldsymbol{Z})] + \overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}. \tag{101}$$

Note that

$$\begin{aligned}
&\sup_n \ \|\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta)^2\Sigma_n(\boldsymbol{Z})]\| \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2|\boldsymbol{Z}]] \\
&\quad = \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2|\boldsymbol{Z}]^2] \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]] \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] < \infty,
\end{aligned} \tag{102}$$

the last step using the eighth moment bound (36). Furthermore,

$$\sup_n \ \|\overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}\| < \infty \tag{103}$$

because $\overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}$ is a convergent sequence by assumption. Hence, $\sup_n \|S_n^2\| < \infty$ and therefore

$$\inf_n \lambda_{\min}(S_n^{-1}) = \inf_n \|S_n\|^{-1} = \inf_n \|S_n^2\|^{-1/2} = \left(\sup_n \|S_n^2\|\right)^{-1/2} > 0.$$

This completes the proof of claim (94).

Finally, we turn to proving claim (95). The claimed convergence of $S_n'^2$ follows immediately from the derivation (99) and the assumption (38). To show that S_n^2 has the same limit, note that the derivation (98) implies that

$$S_n^2 - S_n'^2 = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta_n)^2\Sigma_n(\boldsymbol{Z})] = \frac{1}{n}\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T h_n)^2\Sigma_n(\boldsymbol{Z})].$$

The boundedness of the quantity $\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T h_n)^2\Sigma_n(\boldsymbol{Z})]$ follows by an argument analagous to that in equation (102), which shows that

$$S_n^2 - S_n'^2 \to 0.$$

This completes the proof of statement (95), so we are done. □

C Proofs for Section 7

Proof of Theorem 5. Let us denote

$$[X, \widetilde{X}]_? \equiv (\{X_j, \widetilde{X}_j\}, X_{-j}, \widetilde{X}_{-j}),$$

where $\{X_j, \widetilde{X}_j\}$ represents the *unordered* pair. In other words, $[X, \widetilde{X}]_?$ specifies $[X, \widetilde{X}]$ up to a swap, hence the "?" notation:

$$[X, \widetilde{X}]_? = [x, \widetilde{x}]_? \quad \Longleftrightarrow \quad [X, \widetilde{X}] \in \{[x, \widetilde{x}], [x, \widetilde{x}]_{\mathrm{swap}(j)}\}.$$

With this notation, we claim that

$$T_j^{\mathrm{opt}} \in \arg\max_{T_j} \mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \;\middle|\; [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right] \quad (104)$$

for every $([x, \widetilde{x}], y)$ in the set

$$\mathcal{A} \equiv \left\{([x, \widetilde{x}], y) : T_j^{\mathrm{opt}}([x, \widetilde{x}], y) \neq T_j^{\mathrm{opt}}([x, \widetilde{x}]_{\mathrm{swap}(j)}, y)\right\}. \quad (105)$$

The conclusion (50) will follow because for any T_j,

$$\begin{aligned}
&\mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)] \\
&= \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y), X_j \neq \widetilde{X}_j] \\
&= \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y), ([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{E}\left[\mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| [X, \widetilde{X}]_?, Y\right] \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&\leq \mathbb{E}\left[\mathbb{P}\left[T_j^{\mathrm{opt}}([X, \widetilde{X}], Y) > T_j^{\mathrm{opt}}([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| [X, \widetilde{X}]_?, Y\right] \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{P}\left[T_j^{\mathrm{opt}}([X, \widetilde{X}], Y) > T_j^{\mathrm{opt}}([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)\right].
\end{aligned}$$

The first step holds because $T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)$ implies that $X_j \neq \widetilde{X}_j$, the second by the assumption (49), the third and fourth by probability manipulations, the fifth by the claimed conditional optimality (104), and the sixth by the same logic as the first four steps.

To prove equation (104), fix $([x, \widetilde{x}], y) \in \mathcal{A}$. Consider the simple hypothesis testing problem

$$H_0 : (X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \quad \text{versus} \quad H_1 : (X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j), \quad (106)$$

where $(X_j, \widetilde{X}_j)$ are endowed with their law conditional on

$$([X, \widetilde{X}]_?, Y) = ([x, \widetilde{x}]_?, y).$$

We seek the most powerful test of level $\alpha = 1/2$. Note that under the null distribution, the knockoff exchangeability property makes both events equally likely: $\mathbb{P}_0[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j)] = \mathbb{P}_0[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j)] = 1/2$. Therefore, given any statistic T_j, the level $1/2$ test of the simple hypothesis (106) rejects when $T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y)$. The optimal knockoff statistic T^{opt} defined in equation (104) thus coincides with the most powerful test for the hypothesis (106), which by Neyman-Pearson is given by

$$\begin{aligned}
&T_j^{\text{opt}}([x, \widetilde{x}], y) \\
&\quad = \frac{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right]}{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right]} \\
&\quad = \frac{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?\right] \mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]\right]}{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?\right] \mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]_{\text{swap}(j)}\right]} \\
&\quad = \frac{\mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]\right]}{\mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]_{\text{swap}(j)}\right]} = \frac{\mathbb{P}\left[Y = y \,|X_j = x_j, X_{-j} = x_{-j}\right]}{\mathbb{P}\left[Y = y \,|X_j = \widetilde{x}_j, X_{-j} = x_{-j}\right]}.
\end{aligned}$$

The first step is given by Neyman-Pearson, the second by an application of Bayes rule, the third by the conditional exchangeability of knockoffs (45), and the last by the conditional independence of knockoffs (46). Finally, it is easy to verify that

$$\begin{aligned}
&T_j^{\text{opt}}([X, \widetilde{X}], Y) > T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y) \quad \Longleftrightarrow \\
&\mathbb{P}[Y = y | X_j = x_j, X_{-j} = x_{-j}] > \mathbb{P}[Y = y | X_j = \widetilde{x}_j, X_{-j} = x_{-j}],
\end{aligned}$$

from which we conclude that the likelihood given in equation (48) is optimal for the problem (51). This completes the proof. □

Proof of Proposition 2. Suppose $\boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has a density with respect to the Lebesgue measure. Since

$$\begin{aligned}
&\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] \\
&\quad = \mathbb{E}[\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j} \mid X_{\bullet,-j}, Y, \widetilde{X}]],
\end{aligned}$$

it suffices to show that

$$\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y) \mid X_{\bullet,-j}, Y, \widetilde{X}] = 0$$

for all $X_{\bullet,-j}, Y, \widetilde{X}_j$. Since $\mathcal{L}(\boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}})$ has a density with respect to the Lebesgue measure, so do $\mathcal{L}(\boldsymbol{X}_j | \boldsymbol{Y}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}})$ and $\mathcal{L}(X_j | Y, X_{\bullet,-j}, \widetilde{X})$. Therefore, it suffices to show that the set

$$S(c; x_{\bullet,-j}, y) \equiv \{x_{\bullet,j} : \mathbb{P}(Y = y | X_{\bullet,j} = x_{\bullet,j}, X_{\bullet,-j} = x_{\bullet,-j}) = c\} \subseteq \mathbb{R}^n$$

has Lebesgue measure zero for all $c, x_{\bullet,-j}, y$. To see this, note that if $x_{\bullet,j} \in S(c; x_{\bullet,-j}, y)$, then

$$\begin{aligned} c &= \mathbb{P}(Y = y | X_{\bullet,j} = x_{\bullet,j}, X_{\bullet,-j} = x_{\bullet,-j}) \\ &= \prod_{i=1}^{n} \exp(\eta_i y_i - \psi(\eta_i)) g_0(y_i) \\ &= \exp\left(\sum_{i=1}^{n} (x_{ij}\beta_j + f_{-j}(x_{i,-j})) y_i - \psi(x_{ij}\beta_j + f_{-j}(x_{i,-j})) + \log g_0(y_i) \right). \end{aligned}$$

It follows that

$$\begin{aligned} &S(c; x_{\bullet,-j}, y) \\ &= \left\{ x_{\bullet,j} : \sum_{i=1}^{n} [x_{ij}\beta_j y_i - \psi(x_{ij}\beta_j + f_{-j}(x_{i,-j}))] = \log c - \sum_{i=1}^{n} [f_{-j}(x_{i,-j}) y_i + \log g_0(y_i)] \right\}. \end{aligned} \tag{107}$$

Since ψ is strictly convex and $\beta_j \neq 0$, the left hand side is a strictly concave function of $x_{\bullet,j}$, while the right hand side is a constant (with respect to $x_{\bullet,j}\beta_j$). Thus, $S(c; x_{\bullet,-j}, y)$ is the level set of a strictly concave function, and hence has measure zero. Indeed, the level set of a strictly convex function is the boundary of the corresponding super-level set (which must be convex), and the boundary of any convex set has measure zero [42]. Thus, the conclusion (49) thus follows.

Now, assume that g_η has a density with respect to Lebesgue measure. Since

$$\begin{aligned} &\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] \\ &\quad = \mathbb{E}[\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j} \mid X, \widetilde{X}]], \end{aligned}$$

it suffices to show that

$$\mathbb{P}[P(Y | X_{\bullet,j}, X_{\bullet,-j}) = P(Y | \widetilde{X}_{\bullet,j}, X_{\bullet,-j}) \mid X, \widetilde{X}] = 0 \tag{108}$$

for all $X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}$. From expression (107), we see that $P(Y|X_{\bullet,j}, X_{\bullet,-j}) = P(Y|\widetilde{X}_{\bullet,j}, X_{\bullet,-j})$ if and only if

$$\underbrace{\beta_j (X_{\bullet,j} - \widetilde{X}_{\bullet,j})^T}_{\text{slope}} Y - \underbrace{\psi(\beta_j X_{i,j} + f_{-j}(X_{i,-j})) + \psi(\beta_j \widetilde{X}_{i,j} + f_{-j}(X_{i,-j}))}_{\text{intercept}} = 0.$$

Since $\beta_j \neq 0$ by assumption, the slope $\beta_j(X_{\bullet,j} - \widetilde{X}_{\bullet,j}) \neq 0$ and therefore, the set $\{Y : P(Y|X_{\bullet,j}, X_{\bullet,-j}) = P(Y|\widetilde{X}_{\bullet,j}, X_{\bullet,-j})\}$ is a hyperplane (and hence has Lebesgue measure zero). Together with the fact that Y has a density with respect to Lebesgue measure, this implies the relation (108), so the conclusion (49) follows. □