Electronic Journal of Statistics
Vol. 0 (2022) 1–8
ISSN: 1935-7524

On the power of conditional independence testing under model-X [??]

Eugene Katsevich[??,??] **and Aaditya Ramdas**[??]

Department of Statistics and Data Science, University of Pennsylvania
Department of Statistics and Data Science, Carnegie Mellon University
Machine Learning Department, Carnegie Mellon University
??; **??**

Abstract: For testing conditional independence (CI) of a response Y and a predictor X given covariates Z, the model-X (MX) framework has been the subject of active methodological research, especially in the context of MX knockoffs and their application to genome-wide association studies. In this paper, we study the power of MX CI tests, yielding quantitative insights into the role of machine learning and providing evidence in favor of using likelihood-based statistics in practice. Focusing on the conditional randomization test (CRT), we find that its conditional mode of inference allows us to reformulate it as testing a point null hypothesis involving the conditional distribution of X. The Neyman-Pearson lemma implies that a likelihood-based statistic yields the most powerful CRT against a point alternative. We obtain a related optimality result for MX knockoffs. Switching to an asymptotic framework with arbitrarily growing covariate dimension, we derive an expression for the power of the CRT against local semiparametric alternatives in terms of the prediction error of the machine learning algorithm on which its test statistic is based. Finally, we exhibit a resampling-free test with uniform asymptotic Type-I error control under the assumption that only the first two moments of X given Z are known.

MSC2020 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: sample, LaTeX 2_ε.

Contents

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

[??] arXiv: 2005.05506
[??] Support information of the article.
[??] Some comment
[??] First supporter of the project
[??] Second supporter of the project

of conditional independence (CI),

$$H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z} \quad \text{versus} \quad H_1 : \boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}, \tag{1.1}$$

using n data points

$$(X, Y, Z) \equiv \{(X_i, Y_i, Z_i)\}_{i=1,\dots,n} \overset{\text{i.i.d.}}{\sim} \mathcal{L}. \tag{1.2}$$

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 (**??**). 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 **?**, the problem (**??**) 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. **?**, 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]

$$\textit{model-X (MX) assumption} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}} \text{ for some known } f^*_{\boldsymbol{X}|\boldsymbol{Z}}.^{2} \tag{1.3}$$

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 **?**. On the other hand, the distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ represents the genetic basis of a complex trait,

[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.

[2] We implicitly assume that $\mathcal{L}$ has a density with respect to some dominating measure on $\mathbb{R}^{1+1+p}$, and that all conditional densities are well-defined almost surely. Here and throughout the paper, we identify probability distributions with their densities with respect to the appropriate dominating measure.

about which much less is known. In the context of (stratified) randomized experiments, the distribution $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is the propensity function **?** (the analog of the propensity score for non-binary treatments **?**) 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 **?**. 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 **??** for a brief overview. Rapid progress has been made on the construction of knockoffs in various cases **????** and on the application of this methodology to GWAS **??**. The conditional randomization test (CRT) **?**, 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) **?**, the digital twin test **?**, and the distilled CRT (dCRT) **?**. The dCRT in particular is a promising methodology because it combines good power and computational speed; we focus on this variant of the CRT is Sections **??** and **??** of this paper. We introduce the general CRT methodology next, while deferring the introduction of the dCRT to Section **??**.

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 $\boldsymbol{X}$ (when $\dim(\boldsymbol{X}) = 1$) in a cross-validated lasso **?** of Y on X and Z **?**. To calculate the distribution of T under the null hypothesis (**??**), 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, 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 the upper quantile

$$C(Y,Z) \equiv Q_{1-\alpha}[T(\widetilde{X}, Y, Z)|Y, Z] \tag{1.4}$$

may be extracted (the dependence on α left implicit), where the randomness is over the resampling distribution $\widetilde{X}|Y, Z$. Finally, the CRT rejects if the original test statistic exceeds this quantile:

$$\phi_T^{\mathrm{CRT}}(X,Y,Z) \equiv \begin{cases} 1, & \text{if } T(X,Y,Z) > C(Y,Z); \\ \gamma, & \text{if } T(X,Y,Z) = C(Y,Z); \\ 0, & \text{if } T(X,Y,Z) < C(Y,Z). \end{cases} \tag{1.5}$$

In order to accommodate discreteness, the CRT makes a randomized decision γ when $T(X, Y, Z) = C(Y, Z)$ so that the size of the test is exactly α. In practice, the threshold $C(Y, Z)$ 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, this paper considers only the "infinite-B" version of the CRT as defined by (**??**) and (**??**). 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 **??**, basing inference on rerandomizing the treatment to the units.

The CI testing problem under MX has benefited from several methodological innovations, but fundamental questions regarding power and optimality have received less attention. Therefore, in this paper we address the following two primary questions:

Q1. Are there "optimal" test statistics for MX methods, in any sense?
Q2. What is the precise connection between the performance of the machine learning algorithm and the power of the resulting MX method?

To the best of our knowledge, Q1 has not been considered before, while Q2 has only been indirectly addressed in the context of lasso-based knockoffs **????** and CRT **??**. The present paper complements these existing works by considering arbitrary machine learning methods. We summarize our findings next.

1.3. Our contributions

We find that for the MX CI problem, the CRT is more natural to analyze; it is simpler to analyze than MX knockoffs and is applicable for testing even a single conditional independence hypothesis. Thus, we focus mainly on the CRT in the present paper. We obtain the following nontrivial answers to the questions posed above.

A1: Conditional inference leads to finite-sample optimality against point alternatives. While the composite nonparametric alternative of the CI problem (**??**) 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. Restricting our attention to tests valid conditionally on (Y, Z) (as the CRT is) allows us to reduce the composite null to a point null. We can therefore apply the Neyman-Pearson lemma to show (Section **??**) that the optimal conditionally valid 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:

$$T^{\text{opt}}(X; Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{1.6}$$

The same statistic yields the most powerful one-bit p-values for MX knockoffs (Section **??**). Despite the simplicity of this result, it has not been derived before and appears central to the design of powerful test statistics. Since the model for $Y|X, Z$ is unknown, this result provides our first theoretical indication of the usefulness of machine learning models to learn this distribution (Q2). A2 below gives a more quantitative answer to Q2.

A2: The prediction error of the machine learning method impacts the asymptotic efficiency of the dCRT but not its consistency. It has been widely observed that the better the machine learning 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 dCRT in terms of the prediction error of the underlying machine learning method (Section **??**). In particular, we consider semiparametric alternatives of the form

$$H_1 : \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = N(\boldsymbol{X}^T\beta + g(\boldsymbol{Z}), \sigma^2). \tag{1.7}$$

We analyze the power of a dCRT variant that employs a separately trained estimator $\widehat{g}$ 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 $\mathbb{E}[\text{Var}[\boldsymbol{X}|\boldsymbol{Z}]]$ (denoted s^2). For example, if $d = 1$, the

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

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

On the way to addressing Q2, we additionally establish a third result (Section **??**) that may be of independent interest:

A resampling-free second-order approximation to the dCRT is equivalent to the dCRT and controls Type-I error under weaker assumptions. It was recently pointed out that if $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is Gaussian, then the resampling distribution of the dCRT test statistic can be found in closed form without actual resampling **?**. Here we show that the resampling-free dCRT based on the first two moments of $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is asymptotically equivalent to the dCRT based on $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ itself. Furthermore, we show the former test has asymptotic Type-I error control under the

$$\begin{gathered}\textit{MX(2) assumption:} \text{ the first two moments of } \boldsymbol{X}|\boldsymbol{Z} \text{ are known, i.e.} \\ \mathbb{E}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z}) \text{ and } \text{Var}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \Sigma(\boldsymbol{Z}) \text{ for known } \mu(\cdot), \Sigma(\cdot).\end{gathered} \tag{1.8}$$

This assumption is weaker than the full MX assumption, complementing existing work **??** on weakening assumptions for MX methods. It also suggests that the resampling-free dCRT may be used in place of the usual dCRT while achieving similar power and controlling Type-I error asymptotically.

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 (**??**) and (**??**), population-level variables (such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$) are denoted in boldface, while samples of these variables (such as X_i, Y_i, Z_i) are denoted in regular font. Note that boldface does *not* distinguish between scalars, vectors, and matrices, as it is sometimes employed. The dimensions of the object in this paper will be clear from context. 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 "≡" 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 CRT against point alternatives

In this section, we seek the most powerful CRT against a point alternative. To accomplish this, we make the observation—implicit in earlier works—that the CRT is valid not just unconditionally but also conditionally on Y, Z (Section **??**). 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 **??**). Proofs are deferred to the appendix.

2.1. CRT is conditionally valid and implicitly tests 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 $\boldsymbol{X}|\boldsymbol{Z}$:

$$\begin{aligned} L_0^{\mathrm{MX}}(f^*) &\equiv L_0 \cap L^{\mathrm{MX}}(f^*) \\ &\equiv \{\mathcal{L} : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}\} \cap \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}}\} \\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = f_{\boldsymbol{Z}} \cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}} \cdot f_{\boldsymbol{Y}|\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}, f_{\boldsymbol{Y}|\boldsymbol{Z}}\}. \end{aligned} \tag{2.1}$$

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

$$\sup_{\mathcal{L} \in L_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z)] \le \alpha. \tag{2.2}$$

Recall that the CRT critical value $C(Y, Z)$ is defined via conditional calibration (**??**). As is known to those familiar with MX, this implies that any CRT $\phi = \phi_T^{\mathrm{CRT}}$ not only has level α in the sense of definition (**??**) but also has level α *conditionally* on Y and Z:

$$\sup_{\mathcal{L} \in L_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z) | Y, Z] \le \alpha \quad \text{almost surely.} \tag{2.3}$$

One special property of such conditionally valid tests ϕ is that they can be viewed as testing a *point null* rather than the original *composite null* (**??**). To see this, we view $\phi \equiv \phi(X; Y, Z)$ as a *family* of hypothesis tests, indexed by (Y, Z), for the distribution $\mathcal{L}(X|Y, Z)$. Note that under the MX assumption,

$$\mathcal{L} \in L_0^{\mathrm{MX}}(f^*) \Longrightarrow \mathcal{L}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i). \tag{2.4}$$

In words, fixing Y, Z at their realizations y, z and viewing only X as random, $\mathcal{L}(X|Y = y, Z = z)$ equals a fixed product distribution for any null $\mathcal{L}$. This yields a conditional point null hypothesis, with respect to which $\phi_T^{\mathrm{CRT}}(x; y, z)$ is a level-α test for almost every (y, z). Note that the observations X_i in this conditional distribution are independent *but not identically distributed* due to the different conditioning events in (**??**).

We emphasize that the aforementioned observations have been under the hood of MX papers, and 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. Indeed, viewing the CRT through the conditional lens (**??**) is the starting point that allows us to bring classical theoretical tools to bear on its analysis. We start doing so by considering point alternatives below.

2.2. *The most powerful conditionally valid test*

Viewing the CRT as a test of a point null hypothesis, we can employ the Neyman-Pearson lemma to find the most powerful CRT (in fact, the most powerful conditionally valid test) against point alternatives. The following theorem states that the likelihood ratio with respect to the (unknown) distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ is the most powerful CRT test statistic against a point alternative.

Theorem 2.1. *Let $\bar{\mathcal{L}} \in L^{MX}(f^*)$ 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*

$$T^{\mathrm{opt}}(X, Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{2.5}$$

The CRT $\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}$ based on this test statistic is the most powerful conditionally valid test of $H_0 : \mathcal{L} \in L_0^{\mathrm{MX}}(f^*)$ against $H_1 : \mathcal{L} = \bar{\mathcal{L}}$, i.e.

$$\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X, Y, Z)] \leq \mathbb{E}_{\bar{\mathcal{L}}}[\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}(X, Y, Z)] \tag{2.6}$$

for any test ϕ satisfying the conditional validity property (**??**).

We leave open the question of whether $\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}$ is also the most powerful test among not just conditionally valid tests (**??**) but also among marginally valid tests (**??**). There do at least exist marginally valid tests that are not conditionally valid.

The proof of Theorem **??** (Appendix **??**) is based on the reduction in Section **??** of the composite null to a point null by conditioning, followed by the Neyman-Pearson lemma. Note that the likelihood ratio in the model $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})$ reduces to the likelihood in the model $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$ up to constant factors; see derivation (**??**). This argument has similar flavor to the theory of unbiased testing (see Lehmann and Romano (**?**, 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 **?**, Section 5.10 of Lehmann (1986), now Lehmann and Romano (**?**, Section 5.9)).

Inspecting the most powerful test given by Theorem **??**, 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*

$$\begin{aligned} L_1(f^*,\bar{f}) &= \{\mathcal{L}\in L_0^{\mathrm{MX}}(f^*) : \bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}\} \\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = f_{\boldsymbol{Z}}\cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}}\cdot \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}\}. \end{aligned}$$

Among the set of conditionally valid tests (**??**)*, the test* $\varphi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}$ *is uniformly most powerful against* $L_1(f^*,\bar{f})$*.*

Theorem **??** and Corollary **??** imply that the most powerful CRT against a point alternative is 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

$$\bar{f}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = N(\boldsymbol{X}^T\beta + \boldsymbol{Z}^T\gamma,\sigma^2) \text{ for coefficients } \beta\in\mathbb{R}^d \text{ and } \gamma\in\mathbb{R}^p, \tag{2.7}$$

then the optimal test rejects for small values of $\|Y - X\beta - Z\gamma\|^2$. In Section **??**, 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 **???** using combinations of fitted coefficients $\widehat{\beta}$ and others **??** using likelihood-based test statistics. The above optimality results align more closely with the latter strategy. Theorem **??** has inspired an extension of the CRT to the sequential setting using a likelihood-based test statistic, accompanied by a similar optimality result **?**. Likelihood-based test statistics also have the advantage of avoiding ad hoc combination rules for $\widehat{\beta}\in\mathbb{R}^d$ when $d > 1$. It remains to be seen whether likelihood-based or coefficient-based test statistics yield greater power in practice, but a thorough empirical comparison is beyond the scope of this work. For now, it suffices to note that, despite its simplicity, this is the first such power optimality result in the CRT literature.

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 **??**. We prepare for these results in the next section by exploring an easier-to-analyze asymptotic equivalent to the CRT.

3. An asymptotic equivalent to the distilled CRT

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

3.1. Semiparametric alternatives and the distilled CRT

First, we define an asymptotic framework within which we will work in Sections **??** and **??**. Following a triangular array formalization, for each $n = 1, 2, \dots,$ 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. 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. In this framework, it will be useful to define the mean and variance functions

$$\mu_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \text{ and } \Sigma_n(\boldsymbol{Z}) \equiv \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]. \tag{3.1}$$

Now, consider a set of semiparametric (partially linear) alternatives $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ such that

$$\boldsymbol{Y} = \boldsymbol{X}^T\beta_n + g_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2),\ \sigma^2 > 0 \tag{3.2}$$

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. Of special interest are local alternatives where $\beta_n = h/\sqrt{n}$ for some $h \in \mathbb{R}^d$. We emphasize that—in this section and throughout the paper—we use the partially linear model (**??**) exclusively as an alternative distribution against which to assess power, rather than an additional assumption required for Type-I error control. By Theorem **??**, the most powerful test against the alternative (**??**) is the CRT based on the likelihood statistic

$$\begin{aligned} T_n^{\mathrm{opt}}(X,Y,Z) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - X_i^T\beta_n - g_n(Z_i)\right)^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - (X_i - \mu_n(Z_i))^T\beta_n - g_n'(Z_i)\right)^2\right), \end{aligned} \tag{3.3}$$

where

$$\bar{g}_n(\boldsymbol{Z}) \equiv \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}] = \mu_n(\boldsymbol{Z})^T\beta_n + g_n(\boldsymbol{Z}). \tag{3.4}$$

Assuming local alternatives $\beta_n = h/\sqrt{n}$ and taking a logarithm, we obtain

$$\begin{aligned} \log T_n^{\mathrm{opt}}(X,Y,Z) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (X_i - \mu_n(Z_i))^T h/\sqrt{n} - \bar{g}_n(Z_i)\right)^2 \\ &\approx \frac{h^T}{\sigma^2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \bar{g}_n(Z_i))(X_i - \mu_n(Z_i)) + C, \end{aligned} \tag{3.5}$$

where C is a constant that does not depend on X and therefore does not change upon resampling.

Of course, inference based on T_n^{opt} is infeasible because the function $\bar{g}_n$ is unknown in practice. Suppose we have learned an estimate $\widehat{g}_n$ of this function, possibly in-sample. Then, the derivation (**??**) motivates us to base inference on the sample covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after adjusting for $\boldsymbol{Z}$:

$$\widehat{\rho}_n(X,Y,Z) \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)). \tag{3.6}$$

Consider first the case $d = 1$. The CRT rejecting for large values of $|\widehat{\rho}_n|$ is an instance of the dCRT **?**. The idea of the dCRT (Algorithm **??**) is to *distill*—usually via a machine learning regression method—the information from the high-dimensional $Z \in \mathbb{R}^{p\times n}$ about X and Y into a low-dimensional summary $D \in \mathbb{R}^{q\times n}$, where $q \ll p$. This is accomplished using a *distillation function* $d : (Y,Z) \mapsto D$. Then, the CRT is applied using a test statistic of the form $T_n(X,Y,Z) \equiv T_n^d(X,Y,D) = T_n^d(X,Y,d(Y,Z))$. For example, the CRT based on the statistic $\widehat{\rho}_n$ (**??**) can be expressed as the dCRT with distillation function $d_i(Y,Z) = (\widehat{g}_n(Z_i), \mu_n(Z_i))$, where $\widehat{g}_n$ is learned in-sample on (Y,Z).

[H] **Input:** $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, distribution $f^*_{\boldsymbol{X}|\boldsymbol{Z}}$, distillation function d, test statistic T_n^d, number of resamples B
Distill information in Z about X and Y into $D \equiv d(Y, Z)$ $b = 1, 2, \dots, B$ Resample $\widetilde{X}_i^{(b)} \overset{\text{ind}}{\sim} f^*_{\boldsymbol{X}|\boldsymbol{Z}=Z_i}$, $i = 1, \dots, n$ Compute $\widehat{p} \equiv \frac{1}{B+1} \sum_{b=1}^B \mathbb{1}(T_n^d(\widetilde{X}^{(b)}, Y, D) \geq T_n^d(X, Y, D))$.
Output: dCRT p-value $\widehat{p}$.
Computational cost: One p-dimensional model fit, and drawing B resamples.
The distilled conditional randomization test (dCRT)

The dCRT was proposed for its computational speed: The computationally expensive distillation step is a function only of (Y, Z), so it need not be refit upon resampling $\widetilde{X}$. By contrast, the originally proposed instance of the CRT **?** involved learning $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ on the entire sample (X, Y, Z), and therefore the learning procedure needed to be re-applied to each resampled dataset $(\widetilde{X}^{(b)}, Y, Z)$. The derivations (**??**) and (**??**) suggest that the dCRT is not only computationally fast, but also a natural test to consider for power against semiparametric alternatives (**??**). We therefore focus on this class of tests.

In preparation to study the power of the dCRT, we extend it to $d > 1$ and propose an asymptotically equivalent test that is easier to analyze.

3.2. A second-order approximation to the dCRT

Let us consider first the special case

$$\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\mu_n(\boldsymbol{Z}), \Sigma_n(\boldsymbol{Z})). \tag{3.7}$$

In this case, the resampling distribution of $\widehat{\rho}_n$ can be computed in closed form **?**:

$$\mathcal{L}_n(\sqrt{n} \cdot \widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z) = N(0, \widehat{S}_n^2), \tag{3.8}$$

where

$$\widehat{S}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i). \tag{3.9}$$

When $d = 1$, the dCRT based on the statistic $T_n(X, Y, Z) = |\sqrt{n} \cdot \widehat{\rho}_n(X, Y, Z)|$ (and infinitely many resamples B) therefore rejects when $T_n(X, Y, Z) > \widehat{S}_n \cdot z_{1-\alpha/2}$, requiring no resampling. To extend this to $d > 1$, consider the standardized quantity

$$U_n(X, Y, Z) \equiv \widehat{S}_n^{-1} \sqrt{n} \widehat{\rho}_n = \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \in \mathbb{R}^d. \tag{3.10}$$

It is natural to use as a test statistic the squared norm of U_n:

$$T_n(X, Y, Z) \equiv \|U_n(X, Y, Z)\|^2. \tag{3.11}$$

Then, the normal resampling distribution (**??**) implies that

$$\mathcal{L}_n(T_n(\widetilde{X}, Y, Z) | X, Y, Z) = \chi^2_d. \tag{3.12}$$

It follows that the dCRT based on test statistic $T_n(X, Y, Z)$ yields the test

$$\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > c_{d,1-\alpha}), \tag{3.13}$$

where we recall that $c_{d,1-\alpha}$ is defined as the $1-\alpha$ quantile of χ^2_d. Note that all tests ϕ in Sections **??** and **??** will be (d)CRTs based on the test statistic T_n (**??**). To ease notation, we therefore omit the subscript T_n and the superscript "CRT" from the notation introduced in equation (**??**), replacing these with n and the distribution of $\boldsymbol{X}|\boldsymbol{Z}$ with respect to which resampling is done, respectively. For example, the superscript in the test defined in equation (**??**) is based on the resampling distribution $\boldsymbol{X}|\boldsymbol{Z} \sim N(\mu_n(\boldsymbol{Z}), \Sigma_n(\boldsymbol{Z}))$.

If the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is not Gaussian, then the dCRT $\phi_n^{\mathcal{L}_n}$ based on $T_n(X, Y, Z)$ will not reduce to the closed-form expression (**??**). However, we can think of the test $\phi_n^{N(\mu_n, \Sigma_n)}$ as a kind of second-order approximation for $\phi_n^{\mathcal{L}_n}$ as long as $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ has first and second moments given by $\mu_n(\boldsymbol{Z})$ and $\Sigma_n(\boldsymbol{Z})$, respectively. Indeed, it is easy to check that the resampling distribution $\mathcal{L}_n(\sqrt{n} \cdot \widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z)$ matches that derived in the normal case (**??**) up to two moments. Under a few assumptions, we can make this intuition precise by showing that $\phi_n^{\mathcal{L}_n}$ is asymptotically equivalent to $\phi_n^{N(\mu_n, \Sigma_n)}$ (Theorem **??** below). We require the distribution $\mathcal{L}_n$ to satisfy the following moment conditions[3] for fixed $c_1, c_2 > 0$:

$$\mathcal{L}_n \in L_n(c_1, c_2) \equiv \{\mathcal{L}_n : \|S_n^{-1}\| \le c_1, \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \le c_2\}, \tag{3.14}$$

where

$$S_n^2 \equiv \mathbb{E}[\widehat{S}_n^2] = \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})\right]. \tag{3.15}$$

Furthermore, to avoid technical complications, we assume that the estimate $\widehat{g}_n$ is trained on an independent dataset (whose size can vary arbitrarily with n and is not included in the sample size n used for testing). For example, this independent dataset may be a large observational dataset, while the primary dataset is a smaller experimental one **?**. These training sets across n and resulting estimates $\widehat{g}_n$ remain fixed throughout.

Theorem 3.1. *Suppose that for each n, $\mathcal{L}_n$ is a law whose first and second conditional moments are given by $\mu_n(\boldsymbol{Z})$ and $\Sigma_n(\boldsymbol{Z})$ (**??**), which satisfies the moment conditions (**??**) for fixed for some $c_1, c_2 > 0$. Let $\phi_n^{\mathcal{L}_n}$ be the dCRT based on the test statistic $T_n(X, Y, Z)$ (**??**), (**??**), (**??**), with $\widehat{g}_n$ trained out of sample. The threshold $C_n(Y, Z)$ of this test (**??**) converges in probability to the χ^2_d quantile:*

$$C_n(Y, Z) \overset{\mathcal{L}_n}{\to}_p c_{d,1-\alpha}. \tag{3.16}$$

[3] The exponents in these moment conditions can be relaxed from 4 to $2 + \delta$.

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

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\ \mathbb{P}_{\mathcal{L}_n}[|T_n(X,Y,Z)-c_{d,1-\alpha}|\leq\delta]=0, \tag{3.17}$$

then the dCRT $\phi_n^{\mathcal{L}_n}$ *is asymptotically equivalent to its second order approximation* $\phi_n^{N(\mu_n,\Sigma_n)}$ (**??**)*:*

$$\lim_{n\to\infty}\ \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathcal{L}_n}(X,Y,Z)\neq\phi_n^{N(\mu_n,\Sigma_n)}(X,Y,Z)]=0. \tag{3.18}$$

Informally, this theorem (proved in Appendix **??**) suggests that the CRT resampling distribution of $T_n(X,Y,Z)$ converges to χ^2_d, which is the resampling distribution of this test statistic under a normal $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. Note that the resulting equivalence (**??**) holds for the specific instance of the CRT based on the statistic T_n defined in via equations (**??**) and (**??**), though other kinds of test statistics may lead to similar large-sample behavior. While Theorem **??** is stated for $\widehat{g}_n$ trained out of sample, we conjecture that it continues to hold when $\widehat{g}_n$ is fit in sample, as in the original dCRT construction **?**. At least, we observe that the conditioning in the construction of the resampling distribution $\mathcal{L}_n(\sqrt{n}\cdot\widehat{\rho}_n(\widetilde{X},Y,Z)\mid X,Y,Z)$ ensures that its mean and variance remain equal to 0 and $\widehat{S}_n^2$ even when $\widehat{g}_n$ is fit in sample.

Theorem **??** has several consequences. First, it allows us to study the power of the dCRT $\phi_n^{\mathcal{L}_n}$ against semiparametric alternatives (**??**) by studying instead the simpler test $\phi_n^{N(\mu_n,\Sigma_n)}$. We pursue this direction in Section **??**. Second, it implies a certain robustness property of the dCRT. Indeed, suppose we run the dCRT based on an incorrect law $\mathcal{L}_n'\neq\mathcal{L}_n$, but whose first and second moments match that of $\mathcal{L}_n$ and such that $\mathcal{L}_n$ is contiguous with respect to $\mathcal{L}_n'$. Then, applying Theorem **??** to $\mathcal{L}_n$ and $\mathcal{L}_n'$ implies that $\mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathcal{L}_n'}(X,Y,Z)\neq\phi_n^{\mathcal{L}_n}(X,Y,Z)]\to 0$. It follows that since $\phi_n^{\mathcal{L}_n}$ controls the type-I error asymptotically (in fact, also in finite samples), then so does $\phi_n^{\mathcal{L}_n'}$. We omit the formal statement of this result for the sake of brevity. Third, it suggests a distinct conditional independence test with valid Type-I error control under the weaker assumption that only the first two moments of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ are known. We expand on this third consequence next.

3.3. The MX(2) assumption and the MX(2) F-test

The asymptotic equivalence of $\phi_n^{N(\mu_n,\Sigma_n)}$ to $\phi_n^{\mathcal{L}_n}$ stated in Theorem **??** suggests that we may replace the dCRT based on the law $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ with that based on its normal approximation $N(\mu_n(\boldsymbol{Z}),\Sigma_n(\boldsymbol{Z}))$ while preserving Type-I error control. Since the test $\phi_n^{N(\mu_n,\Sigma_n)}$ requires knowledge only of the first two moments $\mu_n(\boldsymbol{X})$ and $\Sigma_n(\boldsymbol{Z})$, this means that we may control Type-I error without the full MX assumption. To formalize this, let us define the

$$\begin{aligned}&\textit{MX(2) assumption:}\text{ the conditional mean }\mu_n(\boldsymbol{Z})\equiv\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]\\&\text{and conditional variance }\Sigma_n(\boldsymbol{Z})\equiv\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]\text{ are known.}\end{aligned} \tag{3.19}$$

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

$$L_0^{\mathrm{MX}(2)} = L_0^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \equiv L_0 \cap L^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)), \tag{3.20}$$

where

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

Under the MX(2) assumption, the CRT is undefined because there is no conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ to resample from. Nevertheless, we may define the *MX(2) F-test* by running the resampling-free dCRT as though $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ were normal, with the given first and second moments (Algorithm **??**). We denote this test $\phi_n^{N(\mu_n,\Sigma_n)}$, as before.

[H] $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, $\mu_n(\cdot)$ and $\Sigma_n(\cdot)$ in (**??**), learning method g Obtain $\widehat{g}_n$ by fitting g out of sample Recall $\mu_n(Z_i) \equiv \mathbb{E}_{\mathcal{L}_n}[X_i|Z_i]$, set $\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i)$ 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 $T_n = \|U_n\|^2$ Compute $\widehat{p}^{\mathrm{MX}(2)} \equiv \mathbb{P}[\chi_d^2 > T_n]$.
Output: MX(2) F-test asymptotic p-value $\widehat{p}^{\mathrm{MX}(2)}$.
Computational cost: One p-dimensional model fit. **The MX(2) $\boldsymbol{F}$-test**

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 $U_n(X, Y, Z)$.

The MX(2) F-test controls the Type-I error under the MX(2) assumption, if the moment conditions (**??**) hold and $\widehat{g}_n$ is fit out of sample.

Theorem 3.2. *If $\mathcal{L}_n \in L_0^{\mathrm{MX}(2)} \cap L_n(c_1, c_2)$ for some $c_1, c_2 > 0$ and $\widehat{g}_n$ is fit out of sample, then the standardized quantity $U_n(X, Y, Z)$ converges to the standard normal:*

$$U_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{3.21}$$

Therefore, the MX(2) F-test controls Type-I error asympotically, uniformly over the above subset of $L_0^{\mathrm{MX}(2)}$:

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in L_0^{\mathrm{MX}(2)} \cap L_n(c_1,c_2)} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n,\Sigma_n)}(X, Y, Z)] \leq \alpha. \tag{3.22}$$

See Appendix **??** for a proof of this theorem. The moment assumptions can be relaxed for pointwise error control (less desirable), but are unavoidable for uniform type-I error control as stated in the corollary. More importantly, we conjecture that the MX(2) F-test continues to have asymptotic Type-I error control even if $\widehat{g}_n$ is fit in sample. One may expect this because the validity of the MX(2) F-test derives from the correctness of (μ_n, Σ_n) rather than that of $\widehat{g}_n$. This conjecture is supported by the results of a simulation study presented in Appendix **??**.

3.4. Comparison to existing results

Comparison to model-X literature. The preceding results suggest that the MX(2) F-test is a useful alternative to the dCRT: the power of these methods is asymptotically the same (Theorem **??**), while the MX(2) F-test is computationally faster because it does not require resampling (Table **??**). On the other hand, note that we have proven Type-I error control for the MX(2) F-test only when $\widehat{g}_n$ is fit out of sample and only asymptotically, while the dCRT gives finite-sample Type-I error control with in-sample fit $\widehat{g}_n$ (albeit under the stronger model-X assumption). However, numerical simulations suggest good finite-sample Type-I error control for the MX(2) F-test even when $\widehat{g}_n$ is fit in sample. Furthermore, Theorem **??** shows that asymptotic Type-I error control of MX-style methodologies can be achieved under the weaker MX(2) assumption (**??**), requiring only two moments of the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ rather than the entire conditional distribution (Table **??**). If strict Type-I error control in finite samples is desired, however, then we must continue to rely on the full MX assumption. Finally, note that the MX(2) assumption still requires *exact* knowledge of the first and second conditional moments; we leave as an important future direction to examine the robustness of these tests to errors in these quantities. First steps in this direction have been taken recently **??**.

Comparison to doubly-robust literature. The semiparametric model (**??**) has been extensively studied (see e.g. the classic works **??**), in which context estimation of the parameter β_n is well understood. By contrast, we do not assume the validity of the semiparametric model, using it only as an alternative against which to evaluate power. A related and perhaps more relevant line of work is non-parametric doubly robust testing **??** and estimation **??**. Here, the inferential target is some functional of the data-generating distribution. The most relevant such functional is the expected conditional covariance $\rho_n \equiv \mathbb{E}_{\mathcal{L}_n}[\mathrm{Cov}_{\mathcal{L}_n}[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}]]$. Note that a valid test of the null hypothesis $H_0 : \rho_n = 0$ is also a valid test of the conditional independence hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$, since conditional independence implies that $\rho_n = 0$ (though the converse is not true in general). The quantity $\widehat{\rho}_n$ turns out to be a consistent estimator of ρ_n under the MX(2) assumption (Lemma **??**). Such product-of-residuals estimators are also commonly employed in the semi- and non-parametric literatures **???**.

To compare our results with those in non-parametric doubly robust inference, we consider the closest representative of the latter: the generalized covariance measure (GCM) test of Shah and Peters **?**. For $d = 1$, the GCM test statistic is defined as

$$\widehat{\rho}_n^{\mathrm{GCM}} \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \widehat{\mu}_n(Z_i)), \tag{3.23}$$

where $\widehat{\mu}_n(\boldsymbol{Z})$ and $\widehat{g}_n(\boldsymbol{Z})$ are estimates of $\mu_n(\boldsymbol{Z}) = \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $g_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$, respectively. This statistic is shown to converge under conditional independence to a mean-zero normal limit as long as the estimates of $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$

and $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$ are both consistent, while the product in these estimation errors tends to zero at a rate of $o(n^{-1/2})$. By contrast, the MX(2) F-test places more weight on the model for $\boldsymbol{X}|\boldsymbol{Z}$ (assuming both first and second moments of this conditional distribution are known) while placing less weight on the model for $\boldsymbol{Y}|\boldsymbol{Z}$ (not assuming even consistency for $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$). Therefore, while the MX(2) F-test closely resembles the GCM test, the assumptions required for validity of these two methods do not subsume each other (Table **??**).

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

Table 1

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

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

Table 2

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

Comparison to causal inference literature. Theorem **??** is a statement about the asymptotic equivalence between the resampling-based CRT and the asymptotic MX(2) F-test. The MX 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 **???** focusing on the asymptotic behavior of the Fisher randomization test in the context of completely randomized experiments. A similar result to Theorem **??** 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 (**?**, Theorem A.1). Theorem **??** 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.

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

4. Notes

Footnotes[4] pose no problem[5].

5. Displayed text

Text is displayed by indenting it from the left margin. Quotations are commonly displayed. There are short quotations

> This is a short a quotation. It consists of a single paragraph of text. There is no paragraph indentation.

and longer ones.

> This is a longer quotation. It consists of two paragraphs of text. The beginning of each paragraph is indicated by an extra indentation.
>
> This is the second paragraph of the quotation. It is just as dull as the first paragraph.

Another frequently-displayed structure is a list. The following is an example of an *itemized* list, four levels deep.

- This is the first item of an itemized list. Each item in the list is marked with a “tick”. The document style determines what kind of tick mark is used.
- This is the second item of the list. It contains another list nested inside it. The three inner lists are an *itemized* list.
  - This is the first item of an enumerated list that is nested within the itemized list.
  - This is the second item of the inner list. LaTeX allows you to nest lists deeper than you really should.

  This is the rest of the second item of the outer list. It is no more interesting than any other part of the item.
- This is the third item of the list.

The following is an example of an *enumerated* list, four levels deep.

1. This is the first item of an enumerated list. Each item in the list is marked with a “tick”. The document style determines what kind of tick mark is used.
2. This is the second item of the list. It contains another list nested inside it. The three inner lists are an *enumerated* list.
   (a) This is the first item of an enumerated list that is nested within the enumerated list.

   (b) This is the second item of the inner list. LaTeX allows you to nest lists deeper than you really should.

   This is the rest of the second item of the outer list. It is no more interesting than any other part of the item.

[4] This is an example of a footnote.

[5] And another one

Table 3
The spherical case ($I_1 = 0$, $I_2 = 0$).

Equil.					
Points	x	y	z	C	S
L_1	−2.485252241	0.000000000	0.017100631	8.230711648	U
L_2	0.000000000	0.000000000	3.068883732	0.000000000	S
L_3	0.009869059	0.000000000	4.756386544	−0.000057922	U
L_4	0.210589855	0.000000000	−0.007021459	9.440510897	U
L_5	0.455926604	0.000000000	−0.212446624	7.586126667	U
L_6	0.667031314	0.000000000	0.529879957	3.497660052	U
L_7	2.164386674	0.000000000	−0.169308438	6.866562449	U
L_8	0.560414471	0.421735658	−0.093667445	9.241525367	U
L_9	0.560414471	−0.421735658	−0.093667445	9.241525367	U
L_{10}	1.472523232	1.393484549	−0.083801333	6.733436505	U
L_{11}	1.472523232	−1.393484549	−0.083801333	6.733436505	U

3. This is the third item of the list.

The following is an example of a *description* list.

Cow Highly intelligent animal that can produce milk out of grass.
Horse Less intelligent animal renowned for its legs.
Human being Not so intelligent animal that thinks that it can think.

You can even display poetry.

There is an environment for verse
Whose features some poets will curse.
For instead of making
Them do *all* line breaking,
It allows them to put too many words on a line when they'd rather be forced to be terse.

Mathematical formulas may also be displayed. A displayed formula is one-line long; multiline formulas require special formatting instructions.

$$x' + y^2 = z_i^2$$

Don't start a paragraph with a displayed equation, nor make one a paragraph by itself.

Example of a theorem:

Theorem 5.1. *All conjectures are interesting, but some conjectures are more interesting than others.*

Proof. Obvious. □

6. Tables and figures

Cross reference to labelled table: As you can see in Table **??** on page **??** and also in Table **??** on page **??**.

Table 4
Parameter sets used by Bajpai & Reuß

parameter		*Set 1*	*Set 2*
μ_x	$[\mathrm{h}^{-1}]$	0.092	0.11
K_x	[g/g DM]	0.15	0.006
μ_p	[g/g DM h]	0.005	0.004
K_p	[g/L]	0.0002	0.0001
K_i	[g/L]	0.1	0.1
$Y_{x/s}$	[g DM/g]	0.45	0.47
$Y_{p/s}$	[g/g]	0.9	1.2
k_h	$[\mathrm{h}^{-1}]$	0.04	0.01
m_s	[g/g DM h]	0.014	0.029

A major point of difference lies in the value of the specific production rate π for large values of the specific growth rate μ. Already in the early publications **???** it appeared that high glucose concentrations in the production phase are well correlated with a low penicillin yield (the ‘glucose effect’). It has been confirmed recently **????** that high glucose concentrations inhibit the synthesis of the enzymes of the penicillin pathway, but not the actual penicillin biosynthesis. In other words, glucose represses (and not inhibits) the penicillin biosynthesis.

These findings do not contradict the results of **?** and of **?** which were obtained for continuous culture fermentations. Because for high values of the specific growth rate μ it is most likely (as shall be discussed below) that maintenance metabolism occurs, it can be shown that in steady state continuous culture conditions, and with μ described by a Monod kinetics

$$C_s = K_M \frac{\mu/\mu_x}{1-\mu/\mu_x} \tag{6.1}$$

Pirt & Rhigelato determined π for μ between 0.023 and 0.086 h^{-1}. They also reported a value $\mu_x \approx 0.095\ \mathrm{h}^{-1}$, so that for their experiments μ/μ_x is in the range of 0.24 to 0.9. Substituting K_M in Eq. (**??**) by the value $K_M = 1$ g/L as used by **?**, one finds with the above equation $0.3 < C_s < 9$ g/L. This agrees well with the work of **?**, who reported that penicillin biosynthesis repression only occurs at glucose concentrations from $C_s = 10$ g/L on. The conclusion is that the glucose concentrations in the experiments of Pirt & Rhigelato probably were too low for glucose repression to be detected. The experimental data published by Ryu & Hospodka are not detailed sufficiently to permit a similar analysis.

Bajpai & Reuß decided to disregard the differences between time constants for the two regulation mechanisms (glucose repression or inhibition) because of the relatively very long fermentation times, and therefore proposed a Haldane expression for π.

It is interesting that simulations with the **?** model for the initial conditions given by these authors indicate that, when the remaining substrate is fed at a constant rate, a considerable and unrealistic amount of penicillin is produced when the glucose concentration is still very high **???** Simulations with the Bajpai & Reuß model correctly predict almost no penicillin production in similar conditions.

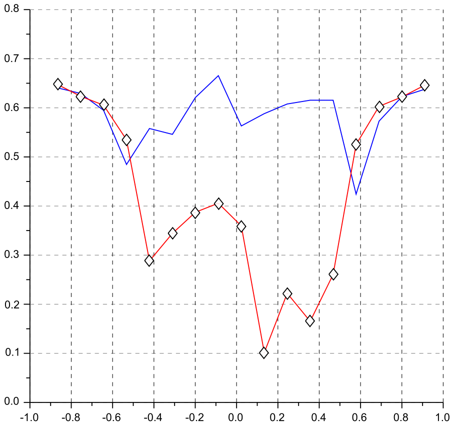

Fig 1. Example of figure inclusion.

Sample of cross-reference to figure. Figure **??** shows that is not easy to get something on paper.

7. Headings

7.1. Subsection

Carr-Goldstein based their model on balancing methods and biochemical knowledge. The original model (1980) contained an equation for the oxygen dynamics which has been omitted in a second paper (1981). This simplified model shall be discussed here.

7.1.1. Subsubsection

Carr-Goldstein based their model on balancing methods and biochemical knowledge. The original model (1980) contained an equation for the oxygen dynamics which has been omitted in a second paper (1981). This simplified model shall be discussed here.

8. Equations and the like

Two equations:

$$C_s = K_M \frac{\mu/\mu_x}{1-\mu/\mu_x} \tag{8.1}$$

and

$$G = \frac{P_{\mathrm{opt}} - P_{\mathrm{ref}}}{P_{\mathrm{ref}}}\ 100\ (\%) \tag{8.2}$$

Two equation arrays:

$$\begin{aligned}
\frac{dS}{dt} &= -\sigma X + s_F F && (8.3)\\
\frac{dX}{dt} &= \mu X && (8.4)\\
\frac{dP}{dt} &= \pi X - k_h P && (8.5)\\
\frac{dV}{dt} &= F && (8.6)
\end{aligned}$$

and,

$$\begin{aligned}
\mu_{\text{substr}} &= \mu_x \frac{C_s}{K_x C_x + C_s} && (8.7)\\
\mu &= \mu_{\text{substr}} - Y_{x/s}(1 - H(C_s))(m_s + \pi / Y_{p/s}) && (8.8)\\
\sigma &= \mu_{\text{substr}} / Y_{x/s} + H(C_s)(m_s + \pi / Y_{p/s}) && (8.9)
\end{aligned}$$

Appendix A: Appendix section

We consider a sequence of queueing systems indexed by n. It is assumed that each system is composed of J stations, indexed by 1 through J, and K customer classes, indexed by 1 through K. Each customer class has a fixed route through the network of stations. Customers in class k, $k = 1, \ldots, K$, arrive to the system according to a renewal process, independently of the arrivals of the other customer classes. These customers move through the network, never visiting a station more than once, until they eventually exit the system.

A.1. Appendix subsection

However, different customer classes may visit stations in different orders; the system is not necessarily "feed-forward." We define the *path of class k customers* in as the sequence of servers they encounter along their way through the network and denote it by

$$\mathcal{P} = \big(j_{k,1}, j_{k,2}, \ldots, j_{k,m(k)}\big). \qquad (A.1)$$

Sample of cross-reference to the formula **??** in Appendix **??**.

Acknowledgments

And this is an acknowledgements section with a heading that was produced by the \section* command. Thank you all for helping me writing this LaTeX sample file.

Supplementary Material

Title of Supplement A

Short description of Supplement A.

Title of Supplement B

Short description of Supplement B.