An Equivalence Test for Conditional Independence Hypotheses

Sydney Kahmann [*]

This version: April 15, 2021

Abstract

Conditional independence relationships serve as the foundation of many observational designs and model assumptions. When evaluating such assumptions, the researcher wishes to falsify their necessary claim, providing evidence in favor of conditional independence. However, current tests are designed to fail to reject, and therefore fail to find evidence against, the intended claim. In this work, I develop an equivalence-based conditional independence test wherein the researcher aims to reject a null of conditional dependence, thus corroborating conditional independence. This test is then applied to causal inference, providing placebo and falsification tests for the conditional ignorability assumption.

[*] skahmann@ucla.edu. Department of Statistics, UCLA. SK was supported by the National Science Foundation Graduate Research Fellowship under Grants No. DGE-1650604 and DGE-2034835.

1 Introduction

Conditional independence relationships are a centerpiece of statistical theory and causality (Dawid, 1979; Rubin, 1974; Spirtes et al., 2000), leveraged by researchers in many applied contexts including genetics (Wille and Bühlmann, 2006), ecology (Denham, Falk, and Mengersen, 2011), and epidemiology (Richardson and Gilks, 1993b). Many methods are based upon assumptions of conditional independence. For example, those in item response theory (IRT), a widely-applicable area of research wherein researchers analyze tests and questionnaires, assuming conditionally independent item responses, to infer latent characteristics about the test and test takers (Lord, 1980). Within the realm of causality, inferences from observational data generally rely upon some assumption of "as-if" randomization, conditional ignorability, or exchangeability, assuming the potential outcomes are conditionally independent of treatment given confounders (Rubin, 1974, Rosenbaum and Rubin, 1983).

Alternatively, practitioners often use conditional independence mappings to infer relationships between observational variables (e.g. Ekici and Onsel, 2013; Mastakouri, Schölkopf, and Janzing, 2019; Morales, Ribas, and Vellido, 2016). These mappings are often graphically displayed. Bayesian networks (Jensen et al., 1996), exponential random graphs (Lusher, Koskinen, and Robins, 2013), and causal directed acyclic graphs (Pearl, 2009), to name a few, all visualize conditional independence and dependence among variables. As a result of the widespread applicability of these graphical interpretations, researchers are actively developing tests of statistical conditional independence for these purposes (e.g. K. Zhang et al., 2012; Doran et al., 2014; Chalupka, Perona, and Eberhardt, 2018; Strobl, K. Zhang, and Visweswaran, 2019).

While conditional independence is assessed in many settings, this work is primarily motivated by applications where the researcher wishes to provide evidence in favor of conditional independence (e.g. when validating model assumptions or observational causal designs). There are two considerations for this context: (1) Conditional independence is an untestable problem, no test for conditional independence uniformly controls Type 1 Error (Shah and Peters, 2018). (2) Hypothesis tests of design are often incorrectly interpreted, providing a failure to find evidence against the claim (Hartman and Hidalgo, 2018).

Regarding the first, as proved in Shah and Peters (2018), no test of conditional independence maintains uniform significance level across the set of alternatives: for each test of conditional independence, there is an edge case the test will not appropriately detect. Therefore, to properly apply a conditional independence test the researcher must understand which alternatives of the test do not have power, a complicated problem in many settings. This work follows Shah and Peters, 2018 in using their regression-based Generalized Covariance Measure (GCM) conditional independence test, which "convert[s] the problem of finding an appropriate test to the more familiar problem of prediction" (p.7). Section 2 reviews the conditional independence testing literature and justifies the choice of GCM as the basis of the equivalence test.

Additionally, the available conditional independence tests are only formulated for tests of difference, i.e. rejecting the null supports a dependence relationship between variables. However, in settings such as causal assumption evaluation, researchers wish to provide support for the null hypothesis that conditional independence does not hold. In existing tests, failing to reject conditional independence cannot be interpreted as acceptance of this null. Although it should be noted: causal assumptions are themselves unverifiable, researchers must evaluate the observable implications of their assumptions, certain tangible properties of the data generating process that should hold if the assumption is true.

To correct these issues with the hypotheses, (2), I propose an equivalence-based test for conditional independence, contributing to the field of statistical equivalence tests (Wellek, 2010; Berger, Hsu, et al., 1996). These equivalence tests are designed to reject a significant deviation from the null as opposed to failing to reject, and improperly accepting, the null outright. In tests of equivalence, the researcher must define the "equivalence range," in this context, the acceptable level of deviation from a perfectly conditionally independent result. As discussed in Section 3, the GCM's covariance-based measure is difficult to credibly specify, thus the equivalence test is developed for use with a correlation-based equivalence range.

The equivalence-based conditional independence test is then applied to problems in causality, primarily, the evaluation of conditional ignorability assumptions. While taking the form of a conditional independence relationship, the inherent missingness of the potential outcomes makes directly evaluating this quantity impossible. Therefore, Section 5.1 proposes a placebo test for the conditional ignorability assumption. If one assumes the "as-if" randomness of treatment, a related placebo outcome should also be "as-if" randomized. Alternatively, Section 5.2 features preliminary work developing a falsification test for this assumption more directly.

2 Background

Conditional independence relationships are a fundamental component of statistics, serving as the foundation of many designs, assumptions, and theorems (Fisher et al., 1920; Fisher, 1934; Fisher et al., 1937; Rosenbaum and Rubin, 1983). In many applications, researchers are interested in conditional independence to learn and communicate the underlying relationships between observational variables (e.g. Richardson and Gilks, 1993a; Richardson and Gilks, 1993b; Mahdi et al., 2012; Mastakouri, Schölkopf, and Janzing, 2019; Ekici and Onsel, 2013; Morales, Ribas, and Vellido, 2016; Wille and Bühlmann, 2006; Denham, Falk, and Mengersen, 2011). Common applications of conditional independence include various model assumptions, such as those in item response theory (IRT), and conditional independence graphs.

With regards to the first, when analyzing tests or questionnaires, researchers often use item response theory to evaluate both the tests and the individuals who take them. IRT methods use test responses to infer the latent characteristics of the test taker, e.g. their unobserved abilities or attitudes, and test, e.g. subclassifications of the various questions. Among other assumptions, IRT methods require local independence, the conditional independence of the test items given the latent variables (Birnbaum, 1968; Lord, 1980). As an example, upholding local independence may require conditioning on test taker proficiency, i.e. a student who answers one question correctly may be more likely to answer another test question correctly (Linden and Glas, 2010). As a result, researchers rely upon various statistics and diagnostics to detect violations of local independence (e.g. Yen, 1984; Chen and Thissen, 1997), thereby testing local dependence. Research in this field is ongoing (e.g. Edwards, Houts, and Cai, 2018; Debelak and Koller, 2020).

Conditional independence relationships are also often expressed via conditional independence graphs or graphical models (Lauritzen, 1996; Jordan et al., 2004). In causality, the causal discovery literature focuses on learning and expressing causal relationships with Bayesian networks and directed acyclic graphs (DAG; Spirtes et al., 2000; Pearl, 2009). In these conditional independence graphs, conditional dependencies are expressed via connections drawn between variables. Each graph is therefore laden with conditional independence assumptions, as the absence of a path between two variables denotes a conditionally independent relationship.

Testing for conditional independence is noted as a hard problem (Bergsma, 2004; Shah and Peters, 2018) and a single graph may require many tests of conditional independence. As an example, recent research attempts to count the potentially vast number of possible DAGs that could result from a partial DAG (e.g. He, Jia, and Yu, 2015; Radhakrishnan, Solus, and Uhler, 2018). Given the difficulty of assessing the web of relationships between a set of covariates, research in conditional independence testing is ongoing (e.g. Fukumizu et al., 2008; Doran et al., 2014; Chalupka, Perona, and Eberhardt, 2018; Shah and Peters, 2018; K. Zhang et al., 2012; Q. Zhang et al., 2017; Strobl, K. Zhang, and Visweswaran, 2019; Heinze-Deml, Peters, and Meinshausen, 2018) and often motivated for these graphical applications. In this section, I formalize statistical conditional independence before introducing these tests.

Statistical conditional independence is often formalized in one of two ways: first, for random variables Y, D, X, D, and X, we say Y is strongly conditionally independent of D given X, formally written as $Y \perp\!\!\!\perp D|X$, if the following conditions hold (Dawid, 1979):

$$p(y, d|x) = p(y|x)p(d|x) \tag{1}$$

$$p(y|d, x) = p(y|x) \tag{2}$$

Alternatively, let $E_1 = \{f \in L^2_{Y,X}, E(f(Y,X)|X) = 0\}$ and $E_2 = \{g \in L^2_{D,X}, E(g(D,X)|X) = 0\}$. For $f \in E_1$ and $g \in E_2$, Y is weakly conditionally independent of D given X, $Y \perp\!\!\!\perp D|X$, if and only if (Daudin, 1980):

$$E(f(Y,X)g(D,X)) = 0. \tag{3}$$

While early conditional independence tests were often built around the definition of strong conditional independence (i.e. Fukumizu et al., 2008), estimation of these distributions was often computationally intensive. Recent research, including the GCM (Shah and Peters, 2018), generally rely upon the weaker definition.

As proved in Shah and Peters (2018), no test of conditional independence maintains uniform significance level across the set of alternatives. Therefore, to properly apply a conditional independence test the researcher must understand which alternatives of the test do not have power. However, as the authors note, even if the researcher understands the circumstances under which a conditional independence test lacks power, they likely do not know the exact data generating process which is often necessary to select an appropriate conditional independence test. This need for an easily-understandable test directly relates to our need of a user-friendly test statistic for purposes of the equivalence range, as will be discussed further in Section 3.2.

As a side note, in the extension I consider unconditional independence testing. These tests are not plagued by the fundamental problems of conditional independence testing described previously, and thus researchers are recommended to use unconditional independence tests instead of conditional independence tests whenever possible (Shah and Peters, 2018). When considering unconditional independence testing in this work I use Hoeffding's D (Hoeffding, 1948), although another test that appropriately maintains level and power could be used.

Conditional independence testing research began with kernel-based tests, which favor tests of strong conditional independence (i.e. Fukumizu et al., 2008), with recent research, including regression-based tests, favoring weak conditional independence (i.e. Shah and Peters, 2018). In the following literature review, I focus on kernel-based and regression-based tests, two of the most common subclassifications of conditional independence tests. For a comprehensive review of the literature, see Li and Fan, 2020.

2.1 Conditional Independence Testing Literature

A primary area of research in conditional independence testing focuses on kernel-based tests such as the Conditional Hilbert-Schmidt Independence Criterion (CHSIC, Fukumizu et al., 2008), Kernel-based Conditional Independence Test (KCIT, K. Zhang et al., 2012), Kernel Conditional Independence Permutation Test (KCIPT, Doran et al., 2014), and approximate kernel-based tests (Strobl, K. Zhang, and Visweswaran, 2019). Using strong conditional independence (Dawid, 1979), kernel mappings capture higher order moments which can provide researchers an option for dealing with the curse of dimensionality. However, estimating kernel matrices is computationally intensive, offsetting some of the accuracy gains.

First, CHSIC maps the data into a reproducing kernel Hilbert space (RKHS) and adapts the Hilbert-Schmidt norm of the normalized cross-covariance operator, introduced as the Hilbert-Schmidt Independence Criterion (HSIC) in Gretton et al. (2005), to discern conditional dependence relationships (Fukumizu et al., 2008). The authors adjusted HSIC to evaluate conditional independence by using the conditional cross-covariance operator, which HSIC (and therefore CHSIC) intuitively is based around: $X \perp\!\!\!\perp Y$ if $P_{XY} = P_X P_Y$. However, the null must be permuted and as a result CHSIC struggles with high-dimensional data (Doran et al., 2014) and larger conditioning sets (K. Zhang et al., 2012), making the test slow for larger datasets as well as less accurate in simulations (Doran et al., 2014; Chalupka, Perona, and Eberhardt, 2018).

KCIT then attempted to address the computational difficulties of estimating the null in CHSIC. In KCIT, another kernel-based test statistic is used with a derived asymptotic distribution under the null (K. Zhang et al., 2012). The null distribution is comparatively easier to compute, it can be approximated by the gamma distribution, but the test statistic requires the eigendecompositions and traces of the kernel matrices. Therefore, the computational gains in estimating the null are largely lost by the required matrix computations.

Meanwhile, KCIPT takes a different approach to this problem by reducing the number of permutations required to estimate the null distribution and employing the maximum mean discrepancy test statistic (MMD) in a two-sample testing problem (Doran et al., 2014). Similarly to CHSIC, KCIPT uses the intuition $X \perp\!\!\!\perp Y | Z$ if $P_{XYZ} = P_{X|Z} P_{Y|Z} P_Z$. Therefore, if we consider our dataset to be drawn from this P_{XYZ} distribution, we wish to permute a sample from P_{XYZ} that simulates this $P_{X|Z} P_{Y|Z} P_Z$ distribution. These two samples can then be compared and should only be similar if the conditional independence relation holds. This test only requires one permutation to estimate the null distribution, however, learning this permutation is somewhat intensive.

Approximations of KCIT and kernel-based methods, such as the Regression-based Conditional Independence Test (RCIT, Strobl, K. Zhang, and Visweswaran, 2019) and Randomized (Conditional) Correlation Test (RCoT, Strobl, K. Zhang, and Visweswaran, 2019), are able to greatly improve the computational efficiency by leveraging Fourier features in approximating KCIT. The latter test, RCoT, is equivalent to the regression-based test from Q. Zhang et al. (2017).

Regression-based tests are another active area of conditional independence testing research. Of these, Hoyer et al. (2009) and Peters et al. (2014) avoid the computational expense of kernels for the added assumptions of the additive noise model (ANM). The REgression with Subsequent Independence Test (RESIT), as introduced by these authors, takes a two-step approach while using ANM assumptions to test for weak conditional independence (Daudin, 1980) of $X \perp\!\!\!\perp Y | Z$ via unconditional independence of the X and Y residuals. In the first step, X and Y are both individually regressed on Z. In the second step, RESIT tests for independence of the residuals of the two models using HSIC. Detecting independence relationships under RESIT is therefore reliant upon the modeling assumptions from step one.

Q. Zhang et al. (2017) introduce KRESIT, a variant of RESIT leveraging kernels, attempting to strike the balance between the comparative speed of RESIT with the lack of modeling assumptions in kernel-based tests. In KRESIT, as opposed to RESIT, the first step regressions are done on RKHS mappings of X and Y, improving the method's ability to detect nonlinear dependencies in the residuals. As noted previously, KRESIT is very similar to RCoT, the two methods differ in how they estimate the null distribution with the latter employing Fourier features.

In Heinze-Deml, Peters, and Meinshausen (2018), among a suite of methods, the authors propose the Residual Prediction Test (RPT). This test scales the residuals from the X on Z regression, evaluating whether the scaled residuals are dependent upon the Y on Z regression. However, this approach requires assumptions on the noise, retaining the correct significance level when the noise is additive. Also, the tests in this paper induce additional requirements on the conditioning variable (i.e. discrete or continuous data), making a combination of these tests required for evaluating conditional ignorability with mixed data types.

The regression-based Generalized Covariance Measure (GCM) (Shah and Peters, 2018) evaluates the covariance between the residuals of the X on Z and Y on Z regressions to evaluate the $X \perp\!\!\!\perp Y | Z$ conditional independence relationship. If the conditional independence relationship holds, the portions of X and Y that remain after conditioning on Z, i.e. the residuals, should be independent and thus have a covariance of zero. The univariate setting requires the conditional expectation functions from each regression are estimated. The multivariate setting requires estimation of the correlation matrix.

In the following sections, I use the GCM as the basis of our equivalence-based conditional independence test. As a regression-based test, the GCM is faster than its kernel-based counterparts and does not sacrifice accuracy. However, the main strength of the GCM is the interpretability of the conditional independence test and hypotheses. The GCM's reliance on covariance, regression models, and residual terms roots the complicated problem of conditional independence testing into fundamental statistical concepts. This foundation is accessible by practitioners with a basic understanding of statistics, and lends itself to the interpretable equivalence-based test introduced in Section 3.1.

2.2 The Generalized Covariance Measure

Recall the weak definition of conditional independence, Equation 3, proposed by Daudin (1980). Thus, for all relationships where $Y \perp\!\!\!\perp D | X$, the $cov(Y, D|X) = 0$, giving rise to the field of covariance-based conditional independence tests. However, recalling the impossibility result in Shah and Peters (2018), observing the $cov(Y, D|X) = 0$ does not ensure $Y \perp\!\!\!\perp D | X$ as there are always alternatives where $cov(Y, D|X) = 0$ and $Y \not\!\perp\!\!\!\perp D | X$. Therefore, this section is structured as follows: after introducing the GCM, I will discuss the GCM's assumptions, providing insights as to when the GCM is a suitable test for conditional independence.

For a given distribution of (Y, D, X), Shah and Peters (2018)'s Generalized Covariance Measure (GCM) rests upon the decomposition of Y and D into portions that depend upon X and remaining noise:

$$Y = f(X) + \epsilon_Y$$
$$D = g(X) + \epsilon_D.$$

After modelling the portions that depend upon X, expressing $f(X) = E(Y|X = x)$ and $g(X) = E(D|X = x)$, the noise in the Y and D compositions can be written as the residual between the observed variable and model:

$$\epsilon = y - f(x)$$
$$\xi = d - g(x).$$

Under weak conditional independence, we should observe a covariance of zero in the portions of Y and D that do not depend upon X, ϵ and ξ, after conditioning on X. Therefore, the GCM test tests the following hypotheses:

$$H_0 : cov(\epsilon\xi|X) = 0 \tag{4}$$
$$H_A : cov(\epsilon\xi|X) \neq 0. \tag{5}$$

By design, $E(\epsilon|X) = 0$ and $E(\xi|X) = 0$, but the $E(\epsilon\xi|X)$ may vary. Thus, deconstructing the conditional covariance of the residuals, the GCM hypotheses can equivalently be expressed as testing $cov(Y, D|X) = 0$ by evaluating $H_0 : E(\epsilon\xi|X) = 0$.

To introduce the GCM test statistic, first define R as the product of the residuals of the predicted models,

$$R = (y - \widehat{f(x)})(d - \widehat{g(x)}) = \hat{\epsilon}\hat{\xi}. \tag{6}$$

In the univariate setting, the GCM test statistic is defined as

$$T^{(n)} = \frac{\sqrt{n}\frac{1}{n}\sum_1^n R_i}{(\frac{1}{n}\sum_1^n R_i^2 - (\frac{1}{n}\sum_1^n R_j)^2)^{1/2}} \tag{7}$$

for observations $i = 1...n$. This test statistic is therefore the expectation of the residual product divided by the standard error of the residual product. Given the GCM assumptions hold, $T^{(n)}$ is distributed standard normal, indicating the conditional independence hypothesis will be rejected with large values of the test statistic.

We now introduce these underlying assumptions of the GCM (Shah and Peters, 2018). The primary assumption requires the product of the mean squared prediction errors (MSPE) to be small. Define the MSPE for a given regression function h as $\text{MSPE}_h = \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - \widehat{h(x_i)})^2$, this first assumption is then:

$$\text{MSPE}_f \text{MSPE}_g = o(n^{-1}), \tag{8}$$

where the MSPEs are with respect to the $Y \sim X$ and $D \sim X$ regressions, respectively. Additionally, assume the average product of the prediction error and expectation of the observed residual is bounded by one, for each respective regression:

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \widehat{f(x_i)})^2 E(\xi^2|X = x) = O(1) \tag{9}$$

$$\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \widehat{g(x_i)})^2 E(\epsilon^2|X = x) = O(1). \tag{10}$$

The GCM also requires the weak assumption of $0 < E(\xi^2\epsilon^2) < \infty$ and that the prediction error is uniformly small. Under these assumptions, the $T^{(n)}$ asymptotically and uniformly converges to a standard normal distribution.

Given these assumptions, the GCM is generally appropriate when the underlying conditional expectation functions are sufficiently fitted to the data, thus shifting this to the common statistical problem of fitting a predictive model. The GCM allows for various model functions, e.g. generalized additive model, gradient boosting, or kernel ridge regression, so that the researcher can select the model that is most appropriate for their dataset, has an appropriately small MSPE product. As with typical modeling problems, subject matter knowledge is helpful in determining the proper model, but most statisticians have experience fitting appropriate predictive models.

3 Equivalence-Based Conditional Independence Test

With an understanding of the mechanics of the GCM, this section inverts the hypotheses to create an equivalence-based test for conditional independence. The GCM as introduced previously may be used to provide support for conditional dependence, i.e. rejecting the null indicates rejecting conditional independence. However in certain settings, such as evaluating model assumptions, the researcher may wish to provide support for conditional independence, i.e. rejecting the null indicates rejecting conditional dependence. Equivalence testing enables this inversion of the null hypotheses.

Canonically, equivalence testing is rooted in biomedical applications, comparing the effectiveness, or "bioequivalence," of two comparable drugs such as the brand name versus generic version of the same medication (Berger, Hsu, et al., 1996). In such settings, regulatory agencies wish to provide evidence that directly supports that the two drugs are equivalent. This requires some predetermined definition of "equivalence" in context. Thus, researchers must define the "equivalence range," the range of values that would indicate an insignificant deviation from the perfectly equivalent result.

Equivalence tests use a null hypothesis of difference so rejecting this hypothesis can be interpreted as finding statistically significant evidence of no difference (Wellek, 2010). Thus, equivalence tests flip the standard null and alternative hypotheses, allowing researchers to test a null inconsistent with their desired claim against the alternative where the ideal hypothesis holds (Hartman and Hidalgo, 2018). This shifts the burden of proof to the researchers, requiring statistically significant evidence to reject the opposite of their claim.

To introduce this key component of the equivalence test, consider the equivalence analogue of the two-sample t-test in which the researcher compares means from the treated and control group, μ_T and μ_C, respectively, after dividing by the standard deviation, σ. The hypotheses for the two-sample equivalence t-test are as follows (Wellek, 2010; Hartman and Hidalgo, 2018):

$$H_0 : \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \text{ or } \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L \tag{11}$$

$$H_A : \epsilon_L \leq \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_U. \tag{12}$$

Defining the equivalence range as (ϵ_L, ϵ_U), these hypotheses test whether the standardized difference in means exceeds the equivalence range. If this difference is within the equivalence range, the researchers conclude the two groups are equivalent.

To develop an equivalence analogue of the GCM, recall the original null hypothesis, Equation 4, is a function of the conditional covariance. Therefore, an equivalence test would require the researcher to bound the equivalence range in terms of covariance, a non-standardized value that is likely difficult to determine in many settings. The equivalence-based GCM is developed in Section 3.1, followed by a correlation-based version of this equivalence test. Given the standardized values and widely understood interpretations of correlation coefficients, I provide guidance in setting this correlation-based equivalence range in Section 3.2 for contexts in which the researcher does not have strong priors to inform determination of the equivalence range. This section is concluded by a replication of the original GCM simulations in Shah and Peters, 2018 and demonstrate the efficacy of the equivalence-based GCM.

3.1 An Interpretable Equivalence-Based Conditional Independence Test

Converting the GCM to an equivalence test requires the interval inclusion method, determining whether the 100(1-2α)% confidence interval on the GCM test statistic contains the chosen equivalence range (Berger, Hsu, et al., 1996). This approach is equivalent to the two one-sided test (TOST, Berger, Hsu, et al., 1996) where two, one-sided t-tests are conducted for both the ϵ_L and ϵ_U, individually. In the standard case, the difference in means is compared to each given epsilon, although a ratio test may also be used (Berger, Hsu, et al., 1996). The TOST is noted to be asymptotically uniformly most powerful (Romano et al., 2005) although with less than optimal power in finite samples (Hartman and Hidalgo, 2018).

Using the intersection union approach, the original GCM hypotheses, Equations 4 and 5, are updated to test whether the conditional covariance exceeds the equivalence range of ϵ_L, ϵ_U:

$$H_0 : cov(\epsilon\xi|X) \geq \epsilon_U \cup cov(\epsilon\xi|X) \leq \epsilon_L \tag{13}$$

$$H_A : \epsilon_L < cov(\epsilon\xi|X) < \epsilon_U. \tag{14}$$

To implement this test, the user must specify equivalence bounds on the conditional covariance $(\epsilon_L, \epsilon_U) = (\sigma^L_{\epsilon\xi|X}, \sigma^U_{\epsilon\xi|X})$.

However, to develop a practical test we must also consider whether researcher's may credibly specify the equivalence range. User-specification of this range in terms of conditional covariance may be challenging in many settings. Lacking strong substantive theory and taking a statistical approach, there is no widely accepted definition of ”negligible” or “low” covariance. If the user does not have a strong prior to inform the value of this term, specification of this term may be difficult to justify. User-specification of the partial correlation, however, a value between negative one and one, would likely be a much simpler task. Additionally, many researchers are familiar with common interpretations of correlation coefficients and may use statistical knowledge to set this value.

Recall the correlation is equivalent to the covariance divided by the standard errors for each term. Therefore, I convert GCM's covariance-based confidence interval to a partial correlation confidence interval before applying the interval inclusion method. I divide $E(\epsilon\xi|X)$ and $SE(\epsilon\xi|X)$ by consistent estimators of $\sigma_{\epsilon|X}$ and $\sigma_{\xi|X}$ to get a consistent estimator in terms of correlation. The correlation-based equivalence hypotheses are thus

$$H_0 : corr(\epsilon\xi|X) \geq \epsilon_U \cup corr(\epsilon\xi|X) \leq \epsilon_L \tag{15}$$

$$H_A : \epsilon_L < corr(\epsilon\xi|X) < \epsilon_U, \tag{16}$$

where the equivalence range is specified in terms of partial correlation, $(\epsilon_L, \epsilon_U) = (\rho^L_{\epsilon\xi|X}, \rho^U_{\epsilon\xi|X})$. The next section offers guidance in setting this correlation-based equivalence range.

3.2 The Equivalence Range

Selecting the equivalence range is a critical task in any implementation of an equivalence test. As Rainey (2014) notes, the size of the equivalence range indicates the strength of the user's claim. In this context, the equivalence range qualifies the definition of a conditionally independent relationship. While an author may argue a given equivalence range indicates a conditionally independent relationship between the two variables, later readers may disagree with that decision.

Recommended practice in equivalence testing is to set the equivalence range using subject matter knowledge (Hartman and Hidalgo, 2018). However, as mentioned previously, practitioners are likely unable to bound the $\sigma_{Y,D|X}$ using substantive knowledge in many contexts. Therefore, the equivalence-based GCM was converted to consider an equivalence range in terms of partial correlation, a value from 0 to 1. This conversion enables practitioners to more credibly state their equivalence range using either substantive knowledge or fundamental principles in statistics.

The statistical literature has several accepted classifications for “negligible” versus “weak” relationships based upon correlations or related measures. While interpretations of correlation coefficients vary, researchers generally agree that $\rho < 0.1$ indicates a very weak or negligible relationship and $0.1 < \rho < 0.3$ indicates a weak relationship (Akoglu, 2018). Alternatively, when discussing correlation with regards to effect size, Cohen (1988) describes a given effect size as low when $\rho < 0.1$ and moderate when $0.1 < \rho < 0.3$.

Combining these two approaches, I propose default values for the correlation-based equivalence range. Using Wellek (2010)'s language for default equivalence range tolerances, define a “strict” tolerance level for this test as $equiv_\rho = 0.1$ and “liberal” tolerance level as $equiv_\rho = 0.3$. In the absence of strong priors on the partial correlation, these defaults may serve as a starting point. However, when subject matter knowledge is available, researchers should select their own equivalence bound instead of relying upon these defaults. I echo Hartman and Hidalgo (2018) in stressing the default settings do not have bias-bounding properties and urge the researcher to justify their chosen equivalence bound, default or not, as a proper definition of “inconsequentially small” variation from a perfectly conditionally independent relationship within the context of their data.

3.3 Simulations

To demonstrate the effectiveness of our equivalence-based GCM, I reproduce the simulations from Shah and Peters, 2018. The authors test $Y \perp\!\!\!\perp D | X$ where $X = N_X$, $D = f_a(X) + N_D$, and $Y = f_a(X) + N_Y$ where $N_X, N_D, N_Y \sim N(0, 1)$ i.i.d. and $f_a(x) = exp(-x^2/2)sin(ax)$. To test the equivalence conversion, replace the N_Y, N_D and correlate the Y and D error terms with a known ρ. In the following simulations, the ϵ_Y, ϵ_D are drawn mean zero with $\sigma_Y, \sigma_D = 1$ and $\sigma_{YD}, \sigma_{DY} = \rho$. Consider the following distributions with 1000 simulations:

1. $X \sim N(0, 1)$, $Y = f_a(X) + 0.3 \cdot \epsilon_Y$, $D = f_a(X) + 0.3 \cdot \epsilon_D$, $a = 2$
2. The same as (1) but with $a = 4$
3. $X_1, X_2 \sim N(0, 1)$ independent, $Y = f_1(X_1) - f_1(X_2) + 0.3 \cdot \epsilon_Y$, $D = f_1(X_1) + f_1(X_2) + 0.3 \cdot \epsilon_D$
4. $X \sim N(0, 1)$, $Y = f_1(X) + 0.3 \cdot \epsilon_Y$, $D = f_1(X) + 0.3 \cdot \epsilon_D$
5. $X \sim N(0, 1)$, $Y = f_2(X) \cdot \epsilon_Y$, $D = f_2(X) \cdot \epsilon_D$

Note, the simulation (4) herein deviates from that in the original paper to consider univariate Y and D instead of bivariate random variables.

Define the equivalence range with the liberal equivalence range of $equiv_\rho = 0.3$. For simulations generated with $\rho < equiv_\rho$, the equivalence-based GCM should reject the null hypothesis of conditional dependence with a decreasing rejection rate as ρ approaches $equiv_\rho$. The test should reject at the α-level at $\rho = equiv_\rho$. After $\rho > equiv_\rho$, the test should fail to reject the conditional dependence relationship.

In Figure 1, the original GCM (red) and equivalence-based GCM (blue) perform as expected. The GCM fails to reject conditional independence when the correlation is close to zero, and the equivalence-based GCM rejects conditional dependence when the correlation is less than the equivalence range. All of the simulations are slightly conservative, consistently rejecting and the equivalence range instead of rejecting at the α-level.

4 Falsification Testing for Causal Design Assumptions

In the past few years, researchers in causal inference and econometrics have pushed for credible causal designs as part of the credibility revolution (Samii, 2016; Angrist and Pischke, 2010). As

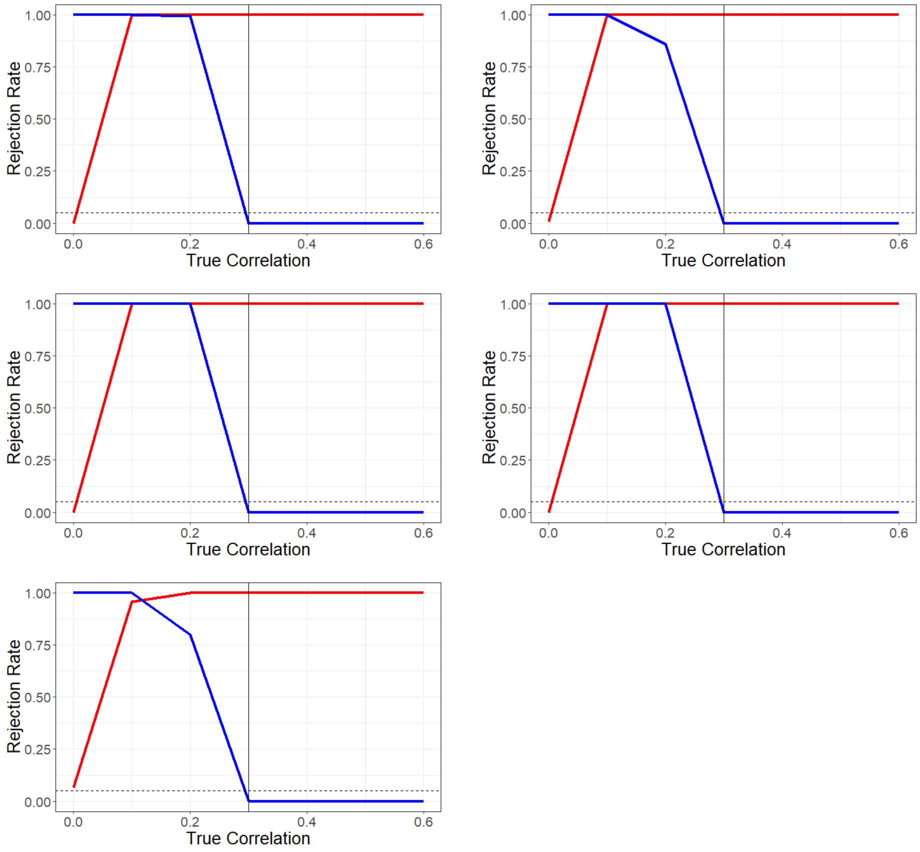

Figure 1: Rejection rates for the GCM test (red) versus equivalence-based GCM test (blue). The GCM test fails to reject conditional independence relationships when the DGP correlation is close to zero. The equivalence-based GCM test rejects a conditionally dependent relationship when the DGP correlation is less than the chosen equivalence range (denoted with vertical line). Simulations 1-5 from left to right, top to bottom.

a result, scientists are increasingly being asked to justify their causal identification assumptions and design-based decisions. This paper seeks to add to the ongoing discussion of best scientific practices by addressing the evaluation of assumptions within the field of observational causal inference.

Equivalence tests are well suited to falsifying causal design assumptions, as typical hypothesis tests would be improperly interpreted in this setting. To falsify a causal assumption, the researcher hopes to provide evidence that the assumption holds by rejecting, or falsifying, a claim inconsistent with their assumption. To avoid interpreting "non-significant difference as significant homogeneity" (Hartman and Hidalgo, 2018), equivalence-based falsification tests invert the null hypotheses of a typical hypothesis test.

As noted previously, causal assumptions themselves are not verifiable, as they make assumptions on the data generating process of the partially-observed set of potential outcomes. Therefore, research in evaluating causal assumptions focuses on evaluating the observable implications of those assumptions.

4.1 Selection on Observables

With experimental data, randomized treatment assignment and the controlled setting often ensures the treatment and control groups are similar in observed and unobserved confounders.[1] Therefore, under a sound experimental design, experimentalists are able to attribute an effect to treatment and not confounding variables.

In the absence of an RCT, the researcher must determine how to construct an appropriate set of controls with which to compare the treatment group, isolating the effects of treatment versus the effects due to confounding variables. Often this is done via an ignorability, exchangeability, or conditional ignorability assumption which ensures treatment assignment is randomly or "as-if" randomly assigned given a set of pre-treatment covariates.

To satisfy "as-if" randomization assumptions, researchers often employ selection on observables designs (SOO, e.g. matching or weighting) to control for confounding variation. These designs artificially adjust the pre-treatment characteristics of the treatment and control groups to appear "comparable." Generally, this comparability is measured via balance, strongly, treatment and control groups have similar distributions across pre-treatment characteristics, or weakly, treatment and control groups have similar means across pre-treatment characteristics. Balance on observables is verifiable, but balance on the potential outcomes is unverifiable.

Matching and weighting designs aim to balance the set of observable covariates between the control and treatment groups. While lacking parametric assumptions, matching methods can run into the curse of dimensionality (Abadie and Imbens, 2006). In high-dimensional settings, observations may not have sufficiently close matches and therefore researchers must remove unmatched data from the analyses. Additionally, as noted in King, Lucas, and Nielsen, 2017, matching methods often face a balance versus matched sample size trade-off: researchers achieve less imbalance with a smaller matched sample size at the cost of higher variance in effect esti-

[1] Although the randomization of treatment assignment in experimental designs theoretically satisfies this unconfoundedness assumption, balance tests can also evaluate whether this condition holds for the given actualization of the randomization scheme.

mates, and vice versa.

Alternatively, Rosenbaum and Rubin, 1983 suggests conditioning on a balancing score, reducing the covariate space to a single variable on which to balance. Traditionally, this is done via propensity score methods. However, propensity scores are parametrically estimated and therefore could result in misspecification bias if estimated incorrectly.

4.2 Conditional Ignorability Assumption

Under the potential outcomes framework (Neyman, 1923, Rubin, 1974), define the set of potential outcomes as $Y_i(0), Y_i(1)$ for individual i where treatment assignment is denoted with $D_i = 1$ for treated units and $D = 0$ for control units. Within an observational setting, researchers assume complete randomization of treatment, the potential outcomes are unconditionally independent of treatment assignment, with the ignorability assumption:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp D_i. \tag{17}$$

This is a strong assumption, thus most observational methods rely upon the conditional ignorability assumption, treatment is randomly assigned given some covariates X. Therefore, the potential outcomes are independent of treatment given covariates X:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp D_i | X. \tag{18}$$

Under this assumption, observed differences in outcomes of interest are due to the only variant, treatment assignment, and not due to distributional differences in observed and unobserved covariates. Using principles of conditional independence, one can equivalently evaluate conditional exogeneity, defined as $D \perp\!\!\!\perp U | X$ where U are unobserved variables.

Under the Structural Causal Model framework, conditional ignorability is defined in terms of relationships, or paths, between causal variables. In a Directed Acyclic Graph (DAG), the lack of a path between two variables is an indication of a conditionally independent relationship between the two. Using this framework, conditional ignorability is satisfied by d-separation, or equivalently, satisfying the backdoor criterion by blocking all backdoor paths between the treatment and outcome variables.

4.3 SOO Evaluation in Practice

Researchers cannot assess conditional ignorability due to the fundamental problem of causal inference where the full set of potential outcomes is unobserved (Holland, 1986). However, researchers can assess balance, an observable implication of this assumption, in two ways: 1) mean balance, comparing the means of observed covariates between groups and 2) distributional balance, comparing the distributions of observed covariates between groups.

For the former, mean balance, the equivalence two sample t-test can be used to compare means between two treatment groups (Wellek, 2010; Hartman and Hidalgo, 2018). This test is straightforward when considering univariate balance, see examples from Hartman and Hidalgo (2018). However, applied researchers often believe several covariates contribute to the data generating process underlying their study and therefore desire mean balance across several covariates. Within this context, multiple testing corrections should be considered.

For the latter, the multivariate case, in lieu of running many low-dimensional balance tests, researchers may employ a single, omnibus balance test (Hansen and Bowers, 2008; Caughey, Dafoe, and Seawright, 2017). These tests have the added benefit of considering joint balance, thereby capturing interactions between covariates.

Along this vein, I propose two extensions for conditional ignorability in Section 5. In Section 5.1, I propose a placebo test of $Y_{t-1} \perp\!\!\!\perp D|X$ for some pre-treatment placebo outcome Y_{t-1}. If the potential outcomes were "as-if" randomly assigned, then one would expect a pre-treatment placebo outcome Y_{t-1} would also be "as-if" randomly assigned. Additionally, in Section 5.2, I provide preliminary work developing a falsification test to assess $Y_0 \perp\!\!\!\perp D|X$ using prognostic and propensity estimates of the Y_0 and D, as this relationship cannot be evaluated directly.

5 Extensions

To assess conditional ignorability, evaluating whether $Y(0), Y(1) \perp\!\!\!\perp D|X$, the observed Y cannot be directly plugged into the equivalence-based conditional independence test: $Y = Y(0)$ when $D = 0$ and $Y = Y(1)$ otherwise, a violation of conditional ignorability when there is an effect of treatment. Additionally, due to the unobserved potential outcomes we cannot use the observed $Y(0)$ and D directly, i.e. $Y(0) \perp\!\!\!\perp D = 0|X$ will always indicate conditional independence. Therefore, I provide two possible extensions, a placebo test and a falsification test for this assumption.

Regarding the placebo test, Section 5.1, if the covariate space captures the confounding between Y and D, and thus conditional ignorability holds, then one might expect a similar, placebo outcome, defined as Y_{t-1}, to also be conditionally independent of D given the covariates.

The falsification test evaluates conditional ignorability more directly in Section 5.2. Propensity and prognostic models may be used to estimate $Y(0)$ and D, respectively. Using these estimates, the falsification tests evaluates the unconditional independence of the two residuals, i.e the observed $Y(0)$ and D compared to their respective models.

Simulations for both extensions use the same data generating process. Generate Y, D as functions of X and their respective error terms, ϵ_Y, ϵ_D. The errors are themselves correlated with a known ρ. See Figure 2 for a graph of the data generating process. As the correlation between ϵ_Y and ϵ_D increases, the conditional independence relationship becomes less believable. With the input of the correlation-based equivalence range, the user specifies at which point the correlation is no longer characterized as representative of a conditionally independent relationship.

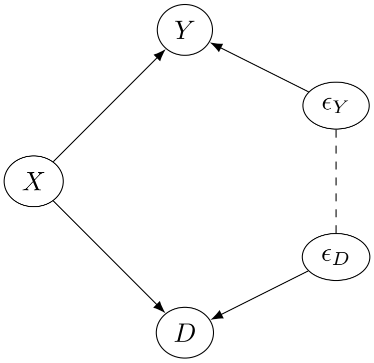

Figure 2: Causal structure for $Y(0) \perp\!\!\!\perp D | X$ where the correlation between the errors is varied.

In the following simulations, set the number of rows as $n = 1000$ and number of simulations as $n.sim = 1000$. The random variables and parameters are defined as follows:

- $X \sim N(0, 1)$
- $Y \sim N(0, 1) + \epsilon_Y$
- $D \sim Binomial(n, p = logit^{-1}(N(0, 1) + \epsilon_D))$
- $(\epsilon_Y, \epsilon_D) \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & corr(\epsilon_Y, \epsilon_D) \\ corr(\epsilon_Y, \epsilon_D) & 1 \end{bmatrix})$

The $corr(\epsilon_Y, \epsilon_D)$ parameter is varied with vector $\rho = sequence(0.1, 0.8, by = 0.05)$.

5.1 Placebo Test Extension

This section uses a placebo outcome, Y_{t-1}, to test $Y_{t-1} \perp\!\!\!\perp D|X$ using the correlation-based approach to the conditional independence equivalence test. Contrary to a potential outcome, the Y_{t-1} is fully observed and therefore this relationship can be tested directly. However, the correlation-based equivalence range must be adjusted to account for the binary treatment assignment.

Given the binomial D, the known $corr(\epsilon_Y, \epsilon_D)$ parameter does not directly translate to the estimate of this term calculated by the equivalence-based GCM. To demonstrate this, I provide Monte Carlo simulations using 5000 simulations. According to the described data generating process, the known correlation in the errors is $\rho = sequence(0.1, 0.8, by = 0.05)$. Therefore, a proper test would estimate a partial correlation close to this value.

In Figure 3, the Monte Carlo simulations compare the observed Y_{t-1} to 1) the inverse logit used to assign treatment, usually unobserved, and 2) the treatment assignment vector, D. In the figure, the first model is unobservable, "propensity" model (the "ground truth") and the latter the observable, "treatment" model. The known correlation parameter is plotted for reference. While practitioners could estimate a propensity model, therefore an "observable" continuous representation of D, this approach would induce undesirable model specification assumptions. Additionally, this approach would fail in scenarios with unobserved confounding such as the described simulations.

To assess whether the original covariance-based GCM test performs as expected, first consider Figure 3 (upper). In this plot, the propensity and treatment assignment models estimate a similar conditional covariance across all levels of the known correlation in the error terms. However, the correlation is consistently, and substantially, underestimated by the binary model in Figure 3 (lower). The propensity score model slightly overestimates the known conditional correlation, with a maximal bias of 5%, for low values of the known correlation.

Consider a likely scenario, the user implements the equivalence-based GCM to test $Y_{t-1} \perp\!\!\!\perp D|X$ with a liberal equivalence tolerance of $\epsilon_\rho = 0.3$, thus considering weak correlation as "equivalent" to conditional independence. Given the bias in the Monte Carlo simulations, the test will return an equivalence result for strong correlations up to 0.75. For smaller chosen equivalence ranges, this gap is present but less severe. At the strict equivalence tolerance of $\epsilon_\rho = 0.1$, the placebo model concludes equivalence for weak correlations of up to 0.20.

To address this issue, let Figure 3 benchmark the discrepancy between the known correla-

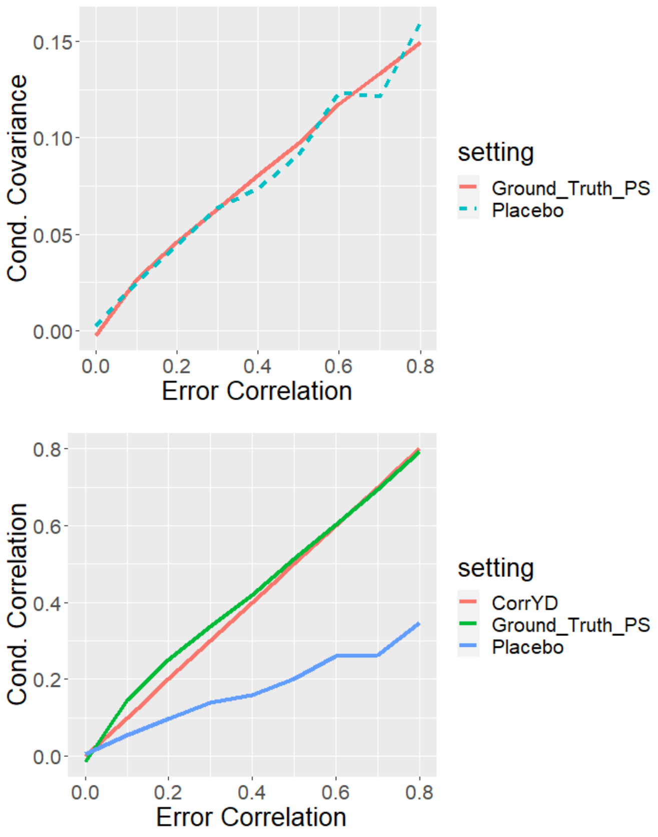

Figure 3: For known correlation between the Y_{t-1} and D errors, consider the Monte Carlo estimates of the conditional covariance (top) and partial correlation (bottom). The partial correlation estimate from the $Y_{t-1}, D|X$ model underestimates the known parameter while the $Y_{t-1}, Propensity|X$ model closely tracks the known partial correlation.

tion and that of the binary D model. For a desired equivalence range allowing weak to negligible levels of correlation, i.e. the liberal range of 0.30, input $equiv_\rho = 0.15$. Using this adjusted equivalence range, rejection rate simulations for the original GCM versus equivalence-based GCM are provided in Figure 4.

According to Figure 4 this method is somewhat conservative. The equivalence test rejects at the α-level at approximately 0.14, slightly before the equivalence range of 0.15. As is ideal, the test consistently rejects the conditional dependence hypothesis for underlying correlations up to 0.05, decreasingly rejects this hypotheses up to a correlation of 0.14, and fails to reject conditional dependence otherwise.

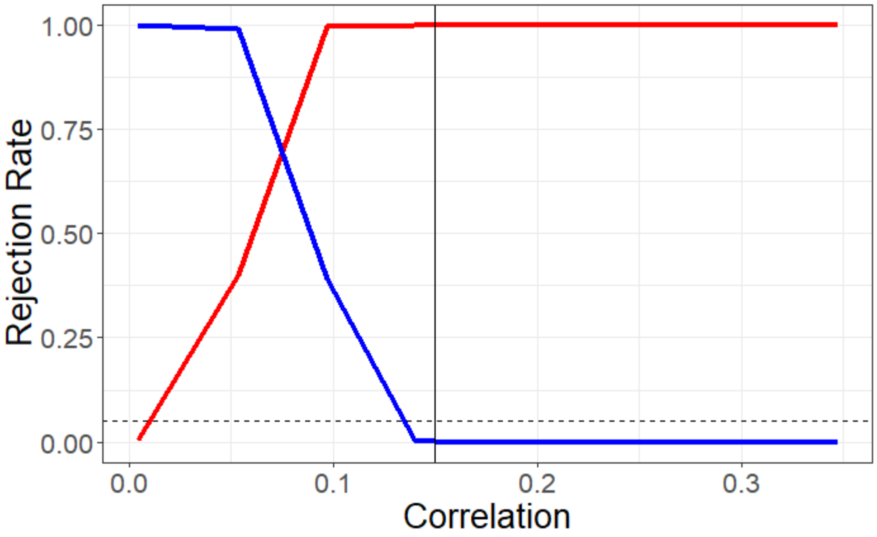

Figure 4: Rejection rate for the original GCM test (red) vs equivalence-based test (blue) of $Y_{t-1}, D|X$. The equivalence range of 0.15 is denoted with vertical line. The equivalence-based test is conservative, rejecting at the α-level just before the equivalence range.

5.2 Falsification Test Extension

The falsification test for conditional ignorability more directly estimates a test of the $Y(0) \perp\!\!\!\perp D|X$ relationship. Recall, $Y(0)$ is not fully observed and thus must be tested only in the control dataset or estimated with a prognostic model fitted using the control dataset. Define the prognostic model of $Y(0)$ with $\Psi(X)$ (Hansen, 2008) and the propensity model of D with $\Phi(X)$ (Rosenbaum and Rubin, 1983), both fitted using the observables, X and lacking the unobservables, ϵ_Y, ϵ_D.

A simple approach to testing the conditional independence of the control potential outcome would be to replace either one or both of $Y(0)$ and D with $\Psi(X)$ and $\Phi(X)$, respectively. As an example, instead of testing $Y(0) \perp\!\!\!\perp D|X$, one could replace $Y(0)$ with the prognostic score $\Psi(X)$

$$\Psi(X) \perp\!\!\!\perp D|X, \tag{19}$$

viewing the prognostic score as an estimate of the control potential outcome, i.e. $\Psi(X) = \widehat{Y(0)}$. However, as would also be true if applied to the propensity score, conditional independence would hold by design when taking $\Psi(X)|X$.

To address this, consider the residual terms, comparing $Y(0)$ and D to the estimated prognostic or propensity model, respectively. Even if the $\Psi(X)$ and $\Phi(X)$ are misspecified, the observed $Y(0)$ and D will contain unobserved confounding. The residuals for the "correctly specified" $Y(0)$ and D versus the potentially incorrectly specified $\Psi(X)$ and $\Phi(X)$, respectively, captures the unobserved confounding and misspecification error. Therefore, in lieu of testing $Y(0) \perp\!\!\!\perp D|X$, we wish to test the following:

$$Y(0) - \Psi(X) \perp\!\!\!\perp D - \Phi(X) \tag{20}$$

. Because X is captured in the propensity and prognostic models, this needs only be evaluated with an unconditional test. In this way, the falsification test evaluates the relationship between the portions of $Y(0$ and D that remain after removing that which depends upon X. [2]

However, again, when $Y(0$ is observed, then $D = 0$. Therefore, the second residual term is the negative propensity score and will not capture unobserved confounding. The correlated Y and D errors in the simulation will not be detected by the conditional independence test. Therefore, the following simulations replace the D in the second residual with the unobserved vector of treatment assignment probabilities, i.e. the second residual is the difference between the true treatment assignment vector and the estimated treatment assignment vector. This updated residual model is compared to the same "ground truth" model as before, i.e testing whether $Y(0)$ is conditionally independent of the unobserved vector of true treatment assignment probabilities given X.

To evaluate whether this residual test, Equation 20, could be a suitable falsification test, Monte Carlo simulations are provided. As shown in Figure 5, the residual test closely follows the ground truth model. However, in many settings the researcher likely does not have access to the unobserved treatment assignment probabilities making this test of questionable practicality.

6 Conclusion

Conditional independence relationships are as a fundamental tenet of statistics with wide-ranging applications. Therefore, this work focuses on developing an equivalence-based conditional independence test. Building upon the Generalized Covariance Measure (GCM) conditional independence test Shah and Peters, 2018, this new approach leverages the interval inclusion method (Berger, Hsu, et al., 1996) to convert the GCM hypotheses into an equivalence framework.

The GCM is well-suited to this purpose, translating the impossible problem of conditional

[2] As noted briefly in Section 2, the unconditional independence test used here is Hoeffding's D (Hoeffding, 1948). Future work would involve discussion of this independence test and formulation of an equivalence version.

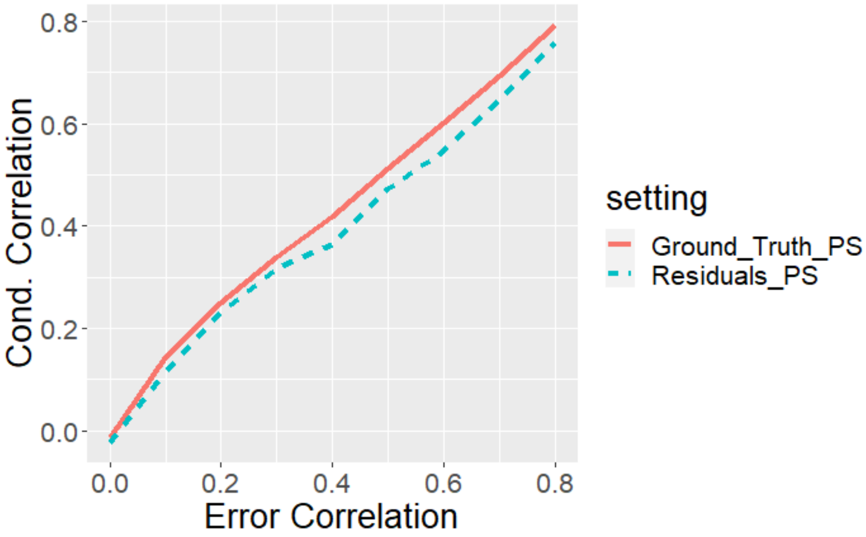

Figure 5: For known correlation between the Y_{t-1} and D errors, consider the Monte Carlo estimates of the partial correlation from the residual models. The propensity-residual model closely tracks the known partial correlation, though the true propensity score is not typically an observed variable.

independence testing (Shah and Peters, 2018) to one in which the user can evaluate the suitability of the method by tackling a standard predictive modeling problem. However, all equivalence tests require user pre-specification of the "equivalence range," the level of deviation from perfect conditional independence that is acceptable within context. Researchers in many fields would likely find GCM's covariance-based measure difficult to pre-specify.

Therefore, the equivalence-based conditional independence test is converted to accept correlation-based equivalence ranges. Practitioners are likely more able to credibly set a correlation-based equivalence range using prior statistical knowledge of negligible or weak correlation as opposed to relying solely on subject matter knowledge to bound a conditional covariance term. In the spirit of Wellek (2010), I propose "strict" and "liberal" default tolerances of this correlation-based equivalence range, $\epsilon_\rho = 0.1$ and $\epsilon_\rho = 0.3$, respectively, based off of common interpretations of correlation and effect size (Cohen, 1988; Akoglu, 2018) for use in cases where the practitioner does not have strong prior beliefs about these values.

This work concludes with two extensions to causal inference, specifically, for evaluation of the conditional ignorability assumption. This problem is complicated by the inherent missingness of potential outcomes, i.e. when $Y(0)$ is observed then $D = 0$. Therefore, evaluating this assumption as it applies to a singular potential outcome, for example $Y(0) \perp\!\!\!\perp D|X$, requires alternative quantities or additional assumptions. The placebo test takes the former approach, while the falsification test of Section 5.2 takes the latter.

The placebo test for conditional ignorability is complicated by the binary treatment assignment. The correlation-based measure of the equivalence-based GCM does not behave as expected when applied to one or more binomial variables. Therefore, Monte Carlo simulations are used to benchmark the adjustment for this discrepancy, allowing the researcher to adjust their desired equivalence range to account for the binomial treatment assignment.

Alternatively, I propose preliminary work on a falsification approach for conditional ignorability. To avoid model misspecification assumptions, consider the residuals of the observed, and thus "fully specified" $Y(0$ and D, as they compare to models of the prognostic and propensity score, respectively. This approach captures instances of misspecification and unobserved confounding but is limited by strong assumptions to avoid the unobservables in causal data.

References

Abadie, Alberto and Guido W Imbens (2006). "Large sample properties of matching estimators for average treatment effects". In: *econometrica* 74.1, pp. 235–267.

Akoglu, Haldun (2018). "User's guide to correlation coefficients". In: *Turkish journal of emergency medicine* 18.3, pp. 91–93.

Angrist, Joshua D and Jörn-Steffen Pischke (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". In: *Journal of economic perspectives* 24.2, pp. 3–30.

Berger, Roger L, Jason C Hsu, et al. (1996). "Bioequivalence trials, intersection-union tests and equivalence confidence sets". In: *Statistical Science* 11.4, pp. 283–319.

Bergsma, Wicher Pieter (2004). *Testing conditional independence for continuous random variables.* Citeseer.

Birnbaum, A Lord (1968). "Some latent trait models and their use in inferring an examinee's ability". In: *Statistical theories of mental test scores.*

Caughey, Devin, Allan Dafoe, and Jason Seawright (2017). "Nonparametric combination (NPC): A framework for testing elaborate theories". In: *The Journal of Politics* 79.2, pp. 688–701.

Chalupka, Krzysztof, Pietro Perona, and Frederick Eberhardt (2018). "Fast Conditional Independence Test for Vector Variables with Large Sample Sizes". In: *arXiv preprint arXiv:1804.02747*.

Chen, Wen-Hung and David Thissen (1997). "Local dependence indexes for item pairs using item response theory". In: *Journal of Educational and Behavioral Statistics* 22.3, pp. 265–289.

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences, 2nd Edition.* Lawrence Erlbaum Associates.

Daudin, JJ (1980). "Partial association measures and an application to qualitative regression". In: *Biometrika* 67.3, pp. 581–590.

Dawid, A Philip (1979). "Conditional independence in statistical theory". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–31.

Debelak, Rudolf and Ingrid Koller (2020). "Testing the local independence assumption of the Rasch model with Q 3-based nonparametric model tests". In: *Applied psychological measurement* 44.2, pp. 103–117.

Denham, Robert J, Matthew G Falk, and Kerrie L Mengersen (2011). "The Bayesian conditional independence model for measurement error: applications in ecology". In: *Environmental and Ecological Statistics* 18.2, pp. 239–255.

Doran, Gary et al. (2014). "A Permutation-Based Kernel Conditional Independence Test." In: *UAI*, pp. 132–141.

Edwards, Michael C, Carrie R Houts, and Li Cai (2018). "A diagnostic procedure to detect departures from local independence in item response theory models." In: *Psychological methods* 23.1, p. 138.

Ekici, Ahmet and Sule Onsel (2013). "How ethical behavior of firms is influenced by the legal and political environments: A Bayesian causal map analysis based on stages of development". In: *Journal of Business Ethics* 115.2, pp. 271–290.

Fisher, Ronald Aylmer et al. (1920). "012: A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error." In:

Fisher, Ronald Aylmer (1934). "Two new properties of mathematical likelihood". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 144.852, pp. 285–307.

Fisher, Ronald Aylmer et al. (1937). "The design of experiments." In: *The design of experiments.* 2nd Ed.

Fukumizu, Kenji et al. (2008). "Kernel measures of conditional dependence". In: *Advances in neural information processing systems*, pp. 489–496.

Gretton, Arthur et al. (2005). "Measuring statistical dependence with Hilbert-Schmidt norms". In: *International conference on algorithmic learning theory.* Springer, pp. 63–77.

Hansen, Ben B (2008). "The prognostic analogue of the propensity score". In: *Biometrika* 95.2, pp. 481–488.

Hansen, Ben B and Jake Bowers (2008). "Covariate balance in simple, stratified and clustered comparative studies". In: *Statistical Science*, pp. 219–236.

Hartman, Erin and F Daniel Hidalgo (2018). "An Equivalence Approach to Balance and Placebo Tests". In: *American Journal of Political Science* 62.4, pp. 1000–1013.

He, Yangbo, Jinzhu Jia, and Bin Yu (2015). "Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs". In: *The Journal of Machine Learning Research* 16.1, pp. 2589–2609.

Heinze-Deml, Christina, Jonas Peters, and Nicolai Meinshausen (2018). "Invariant causal prediction for nonlinear models". In: *Journal of Causal Inference* 6.2.

Hoeffding, Wassily (1948). "A non-parametric test of independence". In: *The annals of mathematical statistics*, pp. 546–557.

Holland, Paul W (1986). “Statistics and causal inference”. In: *Journal of the American statistical Association* 81.396, pp. 945–960.

Hoyer, Patrik O et al. (2009). “Nonlinear causal discovery with additive noise models”. In: *Advances in neural information processing systems*, pp. 689–696.

Jensen, Finn V et al. (1996). *An introduction to Bayesian networks.* Vol. 210. UCL press London.

Jordan, Michael I et al. (2004). “Graphical models”. In: *Statistical science* 19.1, pp. 140–155.

King, Gary, Christopher Lucas, and Richard A Nielsen (2017). “The balance-sample size frontier in matching methods for causal inference”. In: *American Journal of Political Science* 61.2, pp. 473–489.

Lauritzen, Steffen L (1996). *Graphical models.* Vol. 17. Clarendon Press.

Li, Chun and Xiaodan Fan (2020). “On nonparametric conditional independence tests for continuous variables”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 12.3, e1489.

Linden, Wim J van der and Cees AW Glas (2010). “Statistical tests of conditional independence between responses and/or response times on test items”. In: *Psychometrika* 75.1, pp. 120–139.

Lord, Frederic M (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lusher, Dean, Johan Koskinen, and Garry Robins (2013). *Exponential random graph models for social networks: Theory, methods, and applications.* Vol. 35. Cambridge University Press.

Mahdi, Rami et al. (2012). “Empirical Bayes conditional independence graphs for regulatory network recovery”. In: *Bioinformatics* 28.15, pp. 2029–2036.

Mastakouri, Atalanti, Bernhard Schölkopf, and Dominik Janzing (2019). “Selecting causal brain features with a single conditional independence test per feature”. In: *Advances in Neural Information Processing Systems* 32, pp. 12553–12564.

Morales, Carles, Vicent Ribas, and Alfredo Vellido (2016). “Applying conditional independence maps to improve sepsis prognosis”. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW).* IEEE, pp. 254–260.

Neyman, Jerzy S (1923). “On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)”. In: *Annals of Agricultural Sciences* 10, pp. 1–51.

Pearl, Judea (2009). *Causality.* Cambridge university press.

Peters, Jonas et al. (2014). “Causal discovery with continuous additive noise models”. In: *The Journal of Machine Learning Research* 15.1, pp. 2009–2053.

Radhakrishnan, Adityanarayanan, Liam Solus, and Caroline Uhler (2018). “Counting Markov equivalence classes for DAG models on trees”. In: *Discrete Applied Mathematics* 244, pp. 170–185.

Rainey, Carlisle (2014). “Arguing for a negligible effect”. In: *American Journal of Political Science* 58.4, pp. 1083–1091.

Richardson, Sylvia and Walter R Gilks (1993a). “A Bayesian approach to measurement error problems in epidemiology using conditional independence models”. In: *American Journal of Epidemiology* 138.6, pp. 430–442.

— (1993b). “Conditional independence models for epidemiological studies with covariate measurement error”. In: *Statistics in Medicine* 12.18, pp. 1703–1722.

Romano, Joseph P et al. (2005). “Optimal testing of equivalence hypotheses”. In: *The Annals of Statistics* 33.3, pp. 1036–1047.

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

Rubin, Donald B (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of educational Psychology* 66.5, p. 688.

Samii, Cyrus (2016). “Causal empiricism in quantitative research”. In: *The Journal of Politics* 78.3, pp. 941–955.

Shah, Rajen D and Jonas Peters (2018). “The hardness of conditional independence testing and the generalised covariance measure”. In: *arXiv preprint arXiv:1804.07203*.

Spirtes, Peter et al. (2000). *Causation, prediction, and search*. MIT press.

Strobl, Eric V, Kun Zhang, and Shyam Visweswaran (2019). “Approximate kernel-based conditional independence tests for fast non-parametric causal discovery”. In: *Journal of Causal Inference* 7.1.

Wellek, Stefan (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.

Wille, Anja and Peter Bühlmann (2006). “Low-order conditional independence graphs for inferring genetic networks”. In: *Statistical applications in genetics and molecular biology* 5.1, p. 1.

Yen, Wendy M (1984). “Effects of local item dependence on the fit and equating performance of the three-parameter logistic model”. In: *Applied Psychological Measurement* 8.2, pp. 125–145.

Zhang, Kun et al. (2012). “Kernel-based conditional independence test and application in causal discovery”. In: *arXiv preprint arXiv:1202.3775*.

Zhang, Qinyi et al. (2017). “Feature-to-feature regression for a two-step conditional independence test”. In: *Proceedings of the Thirtythird Conference on Uncertainty in Artificial Intelligence*.