A Nonparametric k-Sample Test of Conditional Independence

Nikolas Mittag*
CERGE-EI

December 23, 2018

Abstract: Assumptions that a continuous and a discrete variable are independent conditional on covariates are ubiquitous in, among others, program evaluation, discrete choice models, missing data problems and studies of teacher or peer effects. Instead of testing conditional independence, current studies at best compare means, which only tests correlation. I propose an assumption-free, non-parametric Kolmogorov test that is simple to implement and has power against all alternatives at distance $\sqrt{N}^{-1}$ that differ at any point in the joint support of the distributions of the covariates. Other non- and semi-parametric tests can easily be based on the restriction I test. The test can be used for hypotheses that depend on estimated parameters such as a location shift by (conditional) treatment effects or a regression adjustment. Inference can be conducted by simulation.

Keywords: specification test, conditional independence, balance test
JEL Classification Numbers: C12, C14, C52

*This research was supported the Czech Academy of Sciences (through institutional support RVO 67985998). I would like to thank Stanislav Anatolyev for pointing me in the direction of the solution in this paper, Luca Bittarello, Dan Black, Stéphane Bonhomme, Ivan Canay, Štěpán Jurajda, Jakub Kastl and Azeem Shaikh for comments and Rajeev Dehejia for making the data available. Filip Staněk provided excellent assistance in implementing the test. I am grateful for Bob LaLonde's contributions. All errors are mine. Address: CERGE-EI, joint workplace of Charles University Prague and the Economics Institute of the Academy of Sciences of the Czech Republic, Politických vězňů 7, Prague 1, 111 21, Czech Republic, nikolasmittag@posteo.de

1 Introduction

In this note, I propose a nonparametric, $\sqrt{N}$-consistent test that an outcome y is independent of a discrete variable T conditional on covariates X. The test is universally applicable and simple to implement "out of the box". Many empirical studies assume conditional independence, or equivalently that the distribution of y conditional on X is identical in two or more samples defined by T. For example, program evaluation studies the effect of treatment T on an outcome y using randomized control trials, matching or control functions to separate treatment effects from selection into treatment states. The required assumption that treatment receipt is exogenous conditional on covariates can be examined by testing whether treatment receipt is conditionally independent of variables that are not affected by treatment. Program evaluations also often want to test whether treatment effects are (conditionally) constant or heterogeneous (e.g. Bitler, Gelbach and Hoynes, 2017). This question requires testing whether the conditional distributions of treated and control outcomes are identical after subtracting (conditional) average treatment effects. Similarly, asking whether treatment affects outcomes only through one or multiple channels requires testing whether the distribution of outcomes conditional on X differs by treatment status (Schmieder, von Wachter and Bender, 2016).

More generally, the assumption that y is identically distributed conditional on X in two samples with different marginal distributions of y and X lies at the heart of all re-weighting and control function estimators. These methods are widely applied beyond program evaluation. Testing their validity requires testing whether two conditional distributions are identical after an (estimated) adjustment. For example, methods for missing data usually assume that the data is missing conditionally at random, i.e. that the distribution of the outcome is the same among observations with and without missing data conditional on X (Heitjan and Rubin, 1991). Similarly, the single index assumption of discrete choice models is equivalent to assuming that the covariates y are independent of the discrete dependent variable T conditional on the estimated index X. Decomposing differences between dis-

tributions (Fortin, Lemieux and Firpo, 2011) also raises the question whether conditional distributions are identical.

The two-sample testing problems above are special cases of k-sample tests. Testing whether a discrete variable T that may take more than two values is independent of y conditional on X is a straightforward extension that often arises, for example with multiple treatment states. Common hypotheses that require k-sample tests of conditional independence are studies of peer effects (Elsner and Isphording, 2018) and teacher value added (Chetty, Friedman and Rockoff, 2014). Such studies rely on the assumption that assignment of individuals (teachers or students) to groups (classes) is independent of outcomes (test scores) conditional on the determinants of group assignment. This assumption requires testing whether the distribution of pre-determined predictors of test score growth y are identical across groups defined by T conditional on variables X that determine group assignment.

All examples above make the ubiquitous assumption that $y \perp\!\!\!\perp T|X$. An emerging literature proposes tests comparing *conditional* distributions that are only identical under conditional independence. A few recent papers define test statistics based on the moments (Su and White, 2014; Liu, Wang and Liu, 2018) or non-parametric estimates of these conditional distributions (Su and Spindler, 2013) that allow for both discrete X and T.[1] These tests have known asymptotic distributions. Song (2009) shows a general approach of making tests of conditional independence pivotal. This property comes at the expense of test statistics that depend on estimated functions and transformations, which is inconvenient and may affect finite sample behavior in unknown ways. It also further complicates the test statistics, which depend on (often infinite vectors of) nuisance parameters or require integrating out such infinite vectors or marginal distributions. More importantly, the tests are all based on comparing estimates of *conditional* distribution functions, which makes them depend on

[1]Earlier tests do not apply to the typical applied problems lined out above, because they require a parametric alternative (Andrews, 1997), a parametric model of T (Angrist and Kuersteiner, 2011), all variables in X to be continuous (Delgado and Manteiga, 2001; Su and White, 2007, 2008) or to follow a continuous single-index structure (Song, 2009). Bouezmarni and Taamouti (2014) and Huang, Sun and White (2016) propose tests similar to Su and Spindler (2013) that should be simple to extend to the discrete case.

Kernel estimates. This is impractical, as it implies a difficult bandwidth choice problem (Hansen, 2004) that is amplified by the fact that typical bandwidth choice procedures optimize fit or prediction rather than power or size of the test. Kernel estimates also limit convergence rates of these tests in multidimensional settings (see Huang, Sun and White, 2016, for discussion) and require additional assumptions on differentiability. Only Linton and Gozalo (2014) test conditional independence by comparing a distance measure between empirical cumulative distributions, but their test is for the more general and complicated setting that allows for both variables to be continuous.

The verdict of Imbens and Wooldridge (2009, p. 50) that there are no applications of such methods, still holds, possibly because the available tests are too complicated or infeasible in practice. Applied studies at best compare (conditional) means, which examines whether y and T are (conditionally) uncorrelated rather than conditionally independent.[2] For example, at least 45 out of 107 papers in the 2017 American Economic Review used binary comparisons to identify causal effects (e.g. using randomized assignment, regression discontinuity, difference-in-difference or binary instruments), in which testing for conditional independence of group assignment would be informative. As the exchange between Rothstein (2017) and Chetty, Friedman and Rockoff (2017) on whether teacher assignment to classes is conditionally independent of test score growth underlines, this assumption is crucial for the validity of key results. Yet, none of the papers uses a test of conditional independence.

I propose a simple and general test of conditional independence when T is discrete as in the applications above. It is consistent against all alternatives in which independence fails at some value of X with strictly positive probability in more than one population defined by T and has power against such alternatives at distance $\sqrt{N}^{-1}$. I first show that testing conditional independence for discrete T is equivalent to testing the equality of cumulative distributions. Testing unconditional distributions is a standard problem for which multiple types of tests exist. I propose a non-parametric Kolmogorov test. I describe a bootstrap

[2]Imbens (2015, p.395) discusses testing unconfoundedness by comparing conditional means. Testing for effects on all possible moments or even conditional independence generalizes this strategy.

method that imposes the null hypothesis non-parametrically by estimating the ratio of two joint distributions, but simpler methods can be used if estimating this ratio is infeasible.

Compared to prior tests, this test has three key advantages. First, it is widely applicable, because it is non-parametric and assumption free. Inference remains valid in the presence of estimated parameters, so one can test whether conditional independence holds after a location shift or a regression adjustment. For exposition, I focus on a two-sample test, but my approach trivially extends to k-sample problems with discrete T. Thereby, it can accommodate more than two treatment states and tests for peer effects or teacher value added. Second, the test has power against alternatives at distance $\sqrt{N}^{-1}$ by virtue of being based on the same principle as the test of Linton and Gozalo (2014). The condition I test for discrete T is simpler, of lower dimension and equivalent to conditional independence, which alleviates or solves the problem of low or no power with discrete variables Linton and Gozalo (2014) discuss, simplifies testing, and reduces the computational burden. Third, the test is simple to implement without regard to problem-specific details, because it avoids the data-dependent transformations, nuisance parameters, and the cumbersome choice of bandwidths and other tuning parameters required by the tests discussed above. In addition, the test allows for simple inference via simulation.

Section 2 describes the test statistic, section 3 inference. Section 4 illustrates the test using simulations and data from LaLonde (1986). Section 5 concludes.

2 Test Statistic

Let $F_{y,X|T}(y, X, T = t)$ and $F_{y|X,T}(y, X, T = t)$ denote the joint and conditional distributions of y, X in the two samples $t \in \{0, 1\}$. $\mathcal{X}_t$ is the subset of the d-dimensional sample space of X on which the joint distribution of the covariates $F_{X|T}(X, T = t)$ has strictly positive probability. The data are only informative about differences in the conditional distributions in the joint support $\mathcal{X}_0 \cap \mathcal{X}_1$, so the sample should be restricted to this set. (y, X, T) is the

iid sample of $N = N_0 + N_1$ observations from this common support.[3] $\mathcal{S}_t$ is the subsample for which $T = t$.

Formally, we want to test whether $F_{y|X,T}(y, X, T = 0) = F_{y|X,T}(y, X, T = 1)$, which is equivalent to $\frac{F_{y,X|T}(y,X,T=0)}{F_{X|T}(X,T=0)} = \frac{F_{y,X|T}(y,X,T=1)}{F_{X|T}(X,T=1)}$. This implies the following two equalities

$$F_{y,X|T}(y, X, T = 0) = F_{y,X|T}(y, X, T = 1)\frac{F_{X|T}(X, T = 0)}{F_{X|T}(X, T = 1)} \Leftrightarrow \tag{1}$$

$$F_{y,X|T}(y, X, T = 0)F_{X|T}(X, T = 1) = F_{y,X|T}(y, X, T = 1)F_{X|T}(X, T = 0) \tag{2}$$

The first line illustrates the key idea: the conditional distribution is the same in the two samples if and only if re-weighting the marginal distributions of X so that they are identical also makes the joint distributions of y and X identical. Consequently, one could test whether the empirical joint cumulative distribution from one sample is equal to the reweighted joint distribution of the other sample. Such a test would generalize the common practice of comparing means after re-weighting by the propensity score in that it would test the full assumption of conditional independence and do so entirely non-parametrically, without even imposing a single index structure. The second line suggests a more convenient test statistic:

$$\begin{aligned} CK_{N_0,N_1} = &\sqrt{N_0 \cdot N_1} \max_{j \in \mathcal{S}_0 \cup \mathcal{S}_1} \left| (N_0 \cdot N_1)^{-1} \left(\sum_{i \in \mathcal{S}_0} \mathbb{1}[y_i \leq y_j]\mathbb{1}[X_i \leq X_j] \right) \sum_{i \in \mathcal{S}_1} \mathbb{1}[X_i \leq X_j] \right. \\ &\left. - (N_0 \cdot N_1)^{-1} \left(\sum_{i \in \mathcal{S}_1} \mathbb{1}[y_i \leq y_j]\mathbb{1}[X_i \leq X_j] \right) \sum_{i \in \mathcal{S}_0} \mathbb{1}[X_i \leq X_j] \right| \end{aligned} \tag{3}$$

Where $\mathbb{1}[x \leq X]$ equals 1 if the inequality holds (column-wise) and 0 otherwise. This test statistic is the distance between two empirical cumulative distributions: the joint cumulative distributions of y, X from one sample and an independent draw of X from the marginal distribution of the respective other sample.[4] Using a Kolmogorov-distance makes computation

[3] For ease of exposition, I assume that $\mathcal{X}_0 \cap \mathcal{X}_1$ is a compact set. If this assumption fails, $\mathcal{X}_0 \cap \mathcal{X}_1$ can be partitioned into compact subsets and the same test can be applied with the test statistic being the maximum of the test statistics computed from each compact set.

[4] Alternatively, on can use the maximum Kolmogorov distance of each sample to the pooled sample as the test statistic.

and inference simple, though the statistic differs from a standard Kolmogorov statistic in that it is the supremum over the points in the sample rather than the supremum over all points in the sample space. This makes computation with many covariates feasible, see e.g. Andrews (1997) and Su and White (2014). Choosing other distance metrics, such as the Cramer-von Mises distance, yields similar non-parametric tests.

Theorem 1 *A conditional independence test based on CK_{N_0,N_1} has the following properties:*

1. *The test is consistent against all alternatives in which conditional independence fails to hold at one or more points in the joint support $\mathcal{X}_0 \cap \mathcal{X}_1$.*

2. *The test has power against alternatives at distance $\sqrt{N}^{-1}$ irrespective of the dimension of (y, T, X).*

3. *Critical values of the test can be obtained via the bootstrap without any assumptions.*

Consistency in (1) follows from the fact that (in the joint support) the empirical distributions functions in (3) converge to the functions in equation (2) by definition. Equation (2) holds if and only if conditional independence holds. Property (2) applies to Kolmogorov tests in general (Milbrodt and Strasser, 1990). Property (3), that critical values can be bootstrapped, follows from Romano (1988). The test statistic is the distance between two empirical distributions, so it falls under equation 1.11 of Romano (1988, p.700). Thus, critical values can be bootstrapped without assumptions on the underlying probability law, as long as the maximum is computed over a collection of sets that form a Vapnik-Chervonenkis (VC) class. The test statistic is computed over the set of rectangles defined by the sample points, which is a VC class by virtue of rectangles being convex and having a fixed finite set of extreme points (Dudley, 1978). The theorem extends to k-sample tests using the maximum of the test statistics comparing each sample defined by a specific value of T to the pooled sample.

In many applications, y is a function of estimated parameters, such as when subtracting estimated (conditional) treatment effects or using residuals of a model. One may also sometimes want to impose parametric restrictions on (some of) the distributions in equation

(2). The properties in Theorem 1 continue to hold with estimated parameters under minimal assumptions. For consistency of the test, consistency of the parameter estimates is sufficient. Assumptions 1 (continuous norm differentiability of the parametric model) and 2 (asymptotic linearity and regularity of the estimator) of Romano (1988, example 2) apply to all common estimators and are sufficient for bootstrap tests to still have the correct size.

The test statistic is not sign invariant and depends on the way unordered y are coded.[5] As Andrews (1997) points out, an invariant test can be implemented by using the maximum of these test statistics. One can extend the test in many other ways simply by using the maximum of the test statistic calculated in different ways. For example, one could use the largest test statistic over (all) sign permutations of the columns of X (Andrews, 1997) or over a general class of rectangles (Linton and Gozalo, 2014) to improve power of the test. Peacock (1983) and Fasano and Franceschini (1987) provide systematic generalizations. Methodologically, such extensions amount to taking the maximum in equation (3) over a different collection of sets, so Theorem 1 continues to hold as long as this collection forms a Vapnik-Chervonenkis class that covers the common support.

These extensions of the test may be useful in practice, because a Kolmogorov test inherits both the desirable and the undesirable power properties of such tests. The test converges substantially faster than the Kernel-based tests of Su and White (2007), Su and White (2008), Su and Spindler (2013) or Bouezmarni and Taamouti (2014), which converge at rate $N^{1/2}h^{D/4}$ where h is the bandwidth and D is the dimensionality of (T, X) or even (y, T, X). Tests of the type of Song (2009), Su and Spindler (2013), Huang, Sun and White (2016) or Liu, Wang and Liu (2018) converge faster and can converge at a parametric rate, but to my knowledge, only Linton and Gozalo (2014) establish a convergence rate of order $\sqrt{N}$ that does not depend on the bandwidth and dimensionality. Their test is based on the same principle as the test here, so contrary to all other tests, it is also based on a simple test statistic without estimated nuisance parameters and transformations that affect finite

[5]The distance between joint distributions may not be invariant to changing the sign of (some) variables, so there are 2^d sign permutations of the test statistic for any two samples.

sample behavior in unknown ways. Yet, as all other prior tests that allow for continuous T, test of Linton and Gozalo (2014) is based on the more general implication of conditional independence that $F_{y|T,X} = F_{y|X} \Leftrightarrow F_{y,T,X}F_X = F_{y,X}F_{T,X}$.[6] As Linton and Gozalo (2014) discuss, this more complicated condition of higher dimension likely leads to low power with discrete variables. Thus, for the common case of discrete T, a test based on the simpler condition I propose is likely preferable.

As Janssen (2000) points out, the asymptotic power of any test does not uniformly translate into finite sample power for all samples and alternatives, so which test is preferable in a given application depends on the alternative hypothesis and is thus problem specific. The power function and hence the conditions under which Kolmogorov tests suffer from a lack of power are well known (Milbrodt and Strasser, 1990; Janssen, 1995). For example, the test will be most powerful against alternatives that differ around the median. See Janssen (2000, p. 249) for a general treatment including a discussion of ways to improve finite sample power. If the alternatives of interest indicate low power of the standard test, one could use the well-known refinements of Kolmogorov tests by means of weighting or modifying the maximization problem as discussed above. If power remains low, it may also be useful to conduct tests that have complementary power, for example by using the Cramer-von Mises criterion instead of the Kolmogorov distance to define a test statistic based on equation (3). It would be interesting to verify the conjecture of Su and White (2014) that their test (and the other integrated tests such as Huang, Sun and White 2016 or Liu, Wang and Liu 2018) is complementary. If so, one could further improve matters by conducting such integrated tests based on equation (2) instead of the more general and more complex conditions of the original tests that allow for continuous T.

[6] Theorem 1 of Linton and Gozalo (2014) is very general and covers most Kolmogorov tests as special cases by adapting the distribution functions and rectangles in their test. Their test with $\mathfrak{B}(y) = (-\infty, y), \mathfrak{B}(x) = (-\infty, x), \mathfrak{B}(t) = t$ is equivalent to the alternative test in footnote 4. This test is similar, but not equivalent to the test I propose (among others, it requires more computation). The point of this paper is not to re-invent the general testing principle of Linton and Gozalo (2014), but to show that a much simpler, useful test can be constructed for the common case of discrete T.

3 Inference

An advantage of using a standard statistic such as a Kolmogorov or Cramer-von Mises distance is that inference is well understood. Analytic options include bounding methods (Bierens and Ploberger, 1997), simulating the limiting process (Linton and Gozalo, 2014) or pivotalizing the test statistic (Fasano and Franceschini, 1987; Song, 2009). In finite samples, the bootstrap is often considered preferable (Su and White, 2008; Linton and Gozalo, 2014). In addition to concerns about finite sample performance, using the bootstrap for inference has two key advantages. First, it is easy to implement as an out-of-the-box solution for a wide range of problems. Second, as discussed above, inference remains valid when y is a function of estimated parameters (which need to be re-estimated in every iteration as usual). The general result of Romano (1988) that bootstrap tests have the correct size provides the researcher with many options to obtain critical values via simulation. The answer to the usual question of the bootstrap method with the best finite sample properties depends on the problem at hand, see e.g. MacKinnon (2006) and Horowitz (1997). Simple options that do not impose the null hypothesis are the re-centering approach of Horowitz (1997) or (scalar versions of) the algorithm of Chernozhukov, Fernández-Val and Melly (2013).

For optimal finite sample behavior and to minimize type I errors, it is desirable to impose the null hypothesis on the re-sampling process. Yet, the null hypothesis only specifies a conditional distribution. The marginal distributions of X are left unspecified, but may affect the distribution of the test statistic. To see this, note that the probability limit of the test statistic under the null hypothesis is

$$\sqrt{N_0 N_1} \max_{\mathcal{X}_0 \cup \mathcal{X}_1} \left| \hat{F}_{X|T}(X, T=0) \cdot \hat{F}_{X|T}(X, T=1) \left(\hat{F}_{y|X,T}(y, X, T=1) - \hat{F}_{y|X,T}(y, X, T=0) \right) \right| \quad (4)$$

which depends on the marginal distributions of X, so the bootstrap needs to preserve these distributions. Thus, the usual approach of imposing the null hypothesis by re-sampling from the pooled sample may not work, because it does not preserve $F_{X|T}$.[7] One solution is to re-

[7] Re-sampling from the pooled data approximates the limiting distribution of

sample from $F_{T,X}$ (or F_X) and generate y (or T, y) under the null hypothesis. Su and White (2008) and Su and Spindler (2013) propose such smoothed local bootstrap methods based on Paparoditis and Politis (2000). Linton and Gozalo (2014) take independent draws from the marginal distributions of y and T conditional on X. Both approaches can be applied to the test proposed in this paper, Martin (2007) discusses other methods.

For discrete T, a simple way to preserve the distribution of X and still impose the null hypothesis is to use a weighted bootstrap. In each iteration, one calculates the test statistic from two bootstrap samples of size N_0 and N_1 that are drawn from the pooled sample with different sampling probabilities. To draw the first sample, the sampling probabilities should be $\frac{1-T_i}{N_0+N_1} + T_i \cdot \frac{\hat{f}_{X|T}(x_i,T=0)}{\hat{f}_{X|T}(x_i,T=1)} \cdot \left(\sum_{i\in\mathcal{S}_1} \frac{\hat{f}_{X|T}(x_i,T=0)}{\hat{f}_{X|T}(x_i,T=1)}\right)^{-1} \cdot \frac{N_1}{N_0+N_1}$, where $\hat{f}_{X|T}$ are non-parametric estimates of the distribution of X from each sample. Similarly, the sampling probabilities for the second sample are $\frac{T_i}{N_0+N_1} + (1-T_i) \cdot \frac{\hat{f}_{X|T}(x_i,T=1)}{\hat{f}_{X|T}(x_i,T=0)} \cdot \left(\sum_{i\in\mathcal{S}_1} \frac{\hat{f}_{X|T}(x_i,T=1)}{\hat{f}_{X|T}(x_i,T=0)}\right)^{-1} \cdot \frac{N_0}{N_0+N_1}$. The ratios of the distributions preserve the marginal distributions of X from $\mathcal{S}_0$ and $\mathcal{S}_1$ in the bootstrap samples. The last two terms normalize the probabilities and give weight to the samples according to their original size. Pooling the samples and giving the same aggregate weight to $\mathcal{S}_0$ and $\mathcal{S}_1$ imposes the null hypothesis. See Appendix A for a step-by-step description.

The sampling probabilities depend on the ratio of two estimated multivariate distributions. For inference to be consistent, it is sufficient that the ratio of the estimates converges to the ratio of the true distributions. The denominator is strictly positive, because this distribution is only evaluated at values in the respective sample. Thus, the bootstrap is consistent with any consistent estimator of the distributions. The estimator of Li and Racine (2003) conveniently allows for both continuous and discrete variables. The ideal choice depends on the properties of X, which can be examined empirically in any given case.

The convergence rate of these estimated distributions declines as the number of covariates increases, but contrary to other tests, the convergence rate of the test statistic does

$\sqrt{N_0 N_1} \max_{\mathcal{X}_0\cup\mathcal{X}_1} \left|\hat{F}_X(X)^2 \left(\hat{F}_{y|X}(y,X) - \hat{F}_{y|X}(y,X)\right)\right|$. The second term converges to 0 in both cases, but the difference in the first term may affect the distribution and hence inference. Nevertheless, this equation can sometimes be used for inference (e.g. if $F_{X|T} = F_X$) and adjusting for the differences between $F_{X|T}$ and F_X could simplify inference.

not depend on the number of covariates. Thus, the question is not whether the test has sufficient power in a given sample, but whether the rejection cutoffs are estimated precisely enough. The convergence rate of the bootstrap depends on the convergence rate of the estimated *cumulative* distribution $\hat{F}_{X|T}$, rather than its estimated derivative in the re-sampling probabilities.[8] Thus, the estimated rejection cutoff may converge as slowly as the estimated ratio, but could also converge much faster. The only purpose of the estimated weights is to make the bootstrap account for the impact of differences between the distributions of X in the original samples $\mathcal{S}_0$ and $\mathcal{S}_1$ on equation (4). Whether this is the case can easily be assessed by comparing estimates of $F_{X|T}$ from the bootstrap samples to the original sample. In addition, one can test whether inference distorts the size of the test by simulating y under the null hypothesis either using a parametric model or the Kernel methods referred to above.

Thus, contrary to problems of power, inaccurate inference is easy to detect, so that the weighted bootstrap provides a simple and general way to conduct inference particularly in large samples (relative to the dimension of X). If coverage of the true rejection cutoff is poor, the weighted bootstrap can be refined by adapting the bandwidths or using a different (possibly semi-parametric) estimator for the bootstrap sampling weights. When $F_{X|T}$ is difficult to estimate, either because of small samples or because it has non-standard properties, one can still resort to the other options of inference discussed above.

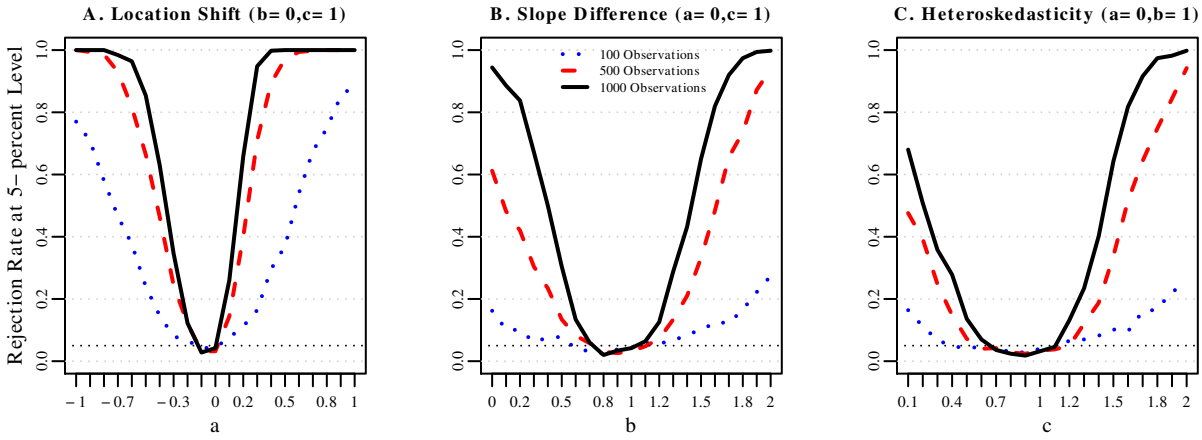
4 Illustration

To illustrate the test in finites samples, I examine how rejection rates vary with sample size and the alternative hypothesis in simulations. See notes of Figure 1 for detail on the simulation setup. To make the marginal distributions differ between samples, I draw X from a standard normal distribution for the first sample and from a bimodal normal mixture for the second sample. The conditional distribution of y given X is a normal distribution with

[8] Note that even estimating the marginal distributions is a problem of the dimension of X, rather than (T, X) or even (y, T, X) as in prior tests.

mean $a + b \cdot x$ and standard deviation c. In the first sample, a is fixed at 0 and both b and c at 1. Figure 1 plots rejection rates at the five-percent level for three types of deviations from the null hypothesis. Panel A varies a, so the conditional distribution is shifted by a constant. Panel B varies b, so the slope of y in X differs. Panel C varies c, so the conditional variance differs. The test slightly under-rejects under the null hypothesis, but its power increases quickly with sample size. The asymmetry in Panel C arises because the difference in the variance of y and hence the common support is larger for low values of c. Note that a test of means cannot detect heteroskedasticity.[9]

Figure 1: Simulated Rejection Rates at the Five Percent Level



X is drawn from a standard normal distribution for the first sample and from a mixture of two normal distributions with mean -0.75 and 0.75, standard deviation 0.8 and 0.5 and mixture probability 0.5 for the second sample. $F_{y|X,T}$ is $\mathcal{N}(x, 1)$ in the first sample and $\mathcal{N}(a + b \cdot x, c)$ in the second sample. The rejection rate at each gridpoint is based on 500 simulations using 1000 bootstrap iterations each. I use the estimator of Li and Racine (2003) for $F_{X|T}$.

To show how the test can be used in practice and that it works well in a multivariate setting, I apply it to the male sample from the National Supported Work Demonstration, which provided temporary employment to a randomized treatment group (LaLonde, 1986;

[9]Comparisons to other tests of conditional independence are complicated, because the tests are difficult to implement and compute highlighting the value of the simplicity of this test. Applying my test to the most closely related simulation setup from the literature, DGP2 of Su and Spindler (2013), yields rejection rates 2-110 percent higher than the rejection rates of their test in table 1 of Su and Spindler (2013). Making the test of Linton and Gozalo (2014) similar to my test as in footnote 6 yields a test that performs similar or worse at about twice the computational cost. Further results are available upon request.

Dehejia and Wahba, 1999, 2002). I use the sample of 445 observations (185 of which were treated) for which wages from two pre-treatment periods (1974 and 1975) are available in addition to the covariates used by LaLonde (age, age squared, Black, years of education, and high school dropout status). This sample allows me to examine the validity of the difference-in-difference estimators as well. The test rejects that pre-treatment (1975) wages are independent of treatment status conditional on the covariates (p-value of 0.098) as well as when also conditioning on 1974 wages (p-value of 0.054). The test does not reject that the distribution of 1974 wages differs by treatment status conditional on the covariates only (p-value of 0.701). These results indicate that randomization has been compromised and that neither the controls nor difference-in-difference methods restore conditional balance. One could use more specific tests to examine how the distributions differ. t-tests of conditional means do not detect a difference in any of the three cases (p-values of 0.340, 0.132 and 0.785).

To examine whether the data support a constant treatment effects model, I test whether the conditional distribution of post-treatment (1978) wages differs between the treatment and control group by more than a constant. The test neither rejects the hypothesis that there is no heterogeneity conditional on the covariates and both 1975 and 1974 wages (p-value of 0.881), nor conditional on 1975 wages only (p-value of 0.792).

5 Conclusion

I propose a non-parametric k-sample test of conditional independence. The test is $\sqrt{N}$-consistent against all alternatives that differ at one or more points in the joint support of X. The test statistic is simple to compute and does not depend on tuning or nuisance parameters. The test remains valid in the presence of estimated parameters. Therefore, the test is widely applicable to problems that require testing whether covariates capture the entire difference between one or more samples. Such problems occur frequently among others in program evaluation, analyses of heterogeneity, discrete choice, missing data or peer

and teacher effects. Simulations indicate that the test works and has power even in small samples. The application shows that it can detect problems that standard tests miss and can thereby provide useful information on the validity of common identifying assumptions.

The key innovation of this paper stems from using the simple condition of equation (2) to test conditional independence for discrete T rather than the more general condition for (mixed) continuous variables prior tests use in combination with defining a test statistic based on unconditional cumulative distributions. These differences yield three main advantages: First, the test is assumption-free, so that it can be applied universally without concerns whether it is appropriate for the nature of the data generating process and the problem at hand. It applies to problems with estimated parameters and easily extends to k-sample problems. Second, it has power against alternatives at distance $N^{-1/2}$ regardless of the dimensionality of the data. Therefore, it is applicable even when samples are small or there are many covariates. Third, the test statistic is simple and does not depend on problem- or data-specific estimates, such as complicated nuisance or tuning parameters. This simplicity makes it easy to implement and apply "out-of-the-box".

Inference is simple to conduct using a weighted bootstrap, but other methods of inference are available if the sample is too small for this convenient option. Even without more specific finite sample results, one can assess size and power of the test in any application using simulations. Such simulations can also be useful to detect cases in which the test has low power against the alternatives of interest (Janssen, 2000). If so, several straightforward extensions and alternatives can be implemented based on the same restriction, such as more powerful (Romano, 1988; Linton and Gozalo, 2014), sign invariant (Andrews, 1997; Fasano and Franceschini, 1987) or pivotal (Justel, Peña and Zamar, 1997) tests. Tests with complementary power properties, such as Cramer-von Mises, integrated (Su and White, 2014; Huang, Sun and White, 2016; Liu, Wang and Liu, 2018) or semi-parametric tests can also be conducted using the simpler restriction of equation (2).

References

Andrews, Donald W. K. 1997. "A Conditional Kolmogorov Test." *Econometrica*, 65(5): 1097–1128.

Angrist, Joshua D., and Guido M. Kuersteiner. 2011. "Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score." *Review of Economics and Statistics*, 93(3): 725–747.

Bierens, Herman J., and Werner Ploberger. 1997. "Asymptotic theory of integrated conditional moment tests." *Econometrica*, 1129–1151.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2017. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *Review of Economics and Statistics*, 99(4).

Bouezmarni, Taoufik, and Abderrahim Taamouti. 2014. "Nonparametric tests for conditional independence using conditional distributions." *Journal of Nonparametric Statistics*, 26(4): 697–719.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on counterfactual distributions." *Econometrica*, 81(6): 2205–2268.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review*, 104(9): 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2017. "Measuring the impacts of teachers: reply." *American Economic Review*, 107(6): 1685–1717.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association*, 94(448): 1053–1062.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity score-matching methods for nonexperimental causal studies." *The review of economics and statistics*, 84(1): 151–161.

Delgado, Miguel A., and Wenceslao G. Manteiga. 2001. "Significance testing in nonparametric regression based on the bootstrap." *Annals of Statistics*, 19(5): 1469–1507.

Dudley, Richard M. 1978. "Central limit theorems for empirical measures." *The Annals of Probability*, 899–929.

Elsner, Benjamin, and Ingo Isphording. 2018. "The Distributional Treatment Problem in the Estimation of Peer Effects." Unpublished Manuscript.

Fasano, Giovanni, and Alberto Franceschini. 1987. "A multidimensional version of the Kolmogorov–Smirnov test." *Monthly Notices of the Royal Astronomical Society*, 225(1): 155–170.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In *Handbook of Labor Economics*. Vol. 4A, Chapter 1, 1–102. Amsterdam:Elsevier.

Hansen, Bruce E. 2004. "Nonparametric conditional density estimation." University of Wisconsin.

Heitjan, Daniel F., and Donald B. Rubin. 1991. "Ignorability and Coarse Data." *The Annals of Statistics*, 19(4): 2244–2253.

Horowitz, Joel L. 1997. "Bootstrap methods in econometrics: theory and numerical performance." In *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress*. Vol. 28 of *Econometric Society Monographs*, , ed. David .M. Kreps and Kenneth F. Wallis, Chapter 7, 188–222. Cambridge:Cambridge University Press.

Huang, Meng, Yixiao Sun, and Halbert White. 2016. "A flexible nonparametric test for conditional independence." *Econometric Theory*, 32(6): 1434–1482.

Imbens, Guido W. 2015. "Matching methods in practice: Three examples." *Journal of Human Resources*, 50(2): 373–419.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47(1): 5–86.

Janssen, Arnold. 1995. "Principal component decomposition of non-parametric tests." *Probability theory and related fields*, 101(2): 193–209.

Janssen, Arnold. 2000. "Global power functions of goodness of fit tests." *Annals of Statistics*, 28(4): 239–253.

Justel, Ana, Daniel Peña, and Rubén Zamar. 1997. "A multivariate Kolmogorov-Smirnov test of goodness of fit." *Statistics & Probability Letters*, 35(3): 251–259.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 76(4): 604–620.

Linton, Oliver, and Pedro Gozalo. 2014. "Testing conditional independence restrictions." *Econometric Reviews*, 33(5-6): 523–552.

Li, Qi, and Jeff Racine. 2003. "Nonparametric estimation of distributions with categorical and continuous data." *Journal of Multivariate Analysis*, 86(2): 266–292.

Liu, Yi, Qihua Wang, and Xiaohui Liu. 2018. "Testing conditional independence via integrating-up transform." *Statistics*, 1–16.

MacKinnon, James. 2006. "Bootstrap Methods in Econometrics." Queen's University Economics Department Working Paper 2-2006.

Martin, Michael A. 2007. "Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties." *Computational Statistics & Data Analysis*, 51(12): 6321–6342.

Milbrodt, Hartmut, and Helmut Strasser. 1990. "On the asymptotic power of the two-sided Kolmogorov-Smirnov test." *Journal of Statistical Planning and Inference*, 26(1): 1–23.

Paparoditis, Efstathios, and Dimitris N Politis. 2000. "The local bootstrap for kernel estimators under general dependence conditions." *Annals of the Institute of Statistical Mathematics*, 52(1): 139–159.

Peacock, JA. 1983. "Two-dimensional goodness-of-fit testing in astronomy." *Monthly Notices of the Royal Astronomical Society*, 202(3): 615–627.

Romano, Joseph P. 1988. "A bootstrap revival of some nonparametric distance tests." *Journal of the American Statistical Association*, 83(403): 698–708.

Rothstein, Jesse. 2017. "Measuring the impacts of teachers: comment." *American Economic Review*, 107(6): 1656–84.

Schmieder, Johannes F., Till von Wachter, and Stefan Bender. 2016. "The effect of unemployment benefits and nonemployment durations on wages." *The American Economic Review*, 106(3): 739–777.

Song, Kyungchul. 2009. "Testing conditional independence via Rosenblatt transforms." *The Annals of Statistics*, 37(6B): 4011–4045.

Su, Liangjun, and Halbert White. 2007. "A consistent characteristic function-based test for conditional independence." *Journal of Econometrics*, 141(2): 807–834.

Su, Liangjun, and Halbert White. 2008. "A nonparametric Hellinger metric test for conditional independence." *Econometric Theory*, 24(4): 829–864.

Su, Liangjun, and Halbert White. 2014. "Testing conditional independence via empirical likelihood." *Journal of Econometrics*, 182(1): 27–44.

Su, Liangjun, and Martin Spindler. 2013. "Nonparametric testing for asymmetric information." *Journal of Business & Economic Statistics*, 31(2): 208–225.

Appendix A: Algorithm for the Weighted Bootstrap

The following procedure implements the weighted bootstrap described in section 3:

1. Compute the test statistic ck from the original samples $\mathcal{S}_0$ and $\mathcal{S}_1$.

2. Calculate $\frac{\hat{f}_{X|T}(x_i,T=1)}{\hat{f}_{X|T}(x_i,T=0)}$ for all $i \in \mathcal{S}_0$ and $\frac{\hat{f}_{X|T}(x_i,T=0)}{\hat{f}_{X|T}(x_i,T=1)}$ for all $i \in \mathcal{S}_1$, where $\hat{f}_{X|T}(x_i, T = t), t = \{0, 1\}$ are non-parametric estimates of the distribution of X from each sample.

3. Repeat the following steps many times:

   (a) Draw a sample $\mathcal{BS}_0$ of size N_0 with replacement from the pooled sample $\{\mathcal{S}_0, \mathcal{S}_1\}$ with the probability of sampling a specific x_i being

$$\Pr(x_i \in \mathcal{BS}_0) = \begin{cases} \dfrac{1}{N_0 + N_1} & \text{if } i \in \mathcal{S}_0 \\ \dfrac{\hat{f}_{X|T}(x_i, T = 0)}{\hat{f}_{X|T}(x_i, T = 1)} \cdot \dfrac{1}{\sum_{i \in \mathcal{S}_1} \frac{\hat{f}_{X|T}(x_i,T=0)}{\hat{f}_{X|T}(x_i,T=1)}} \cdot \dfrac{N_1}{N_0 + N_1} & \text{if } i \in \mathcal{S}_1 \end{cases} \tag{5}$$

   (b) Similarly, draw a sample $\mathcal{BS}_1$ of size N_1 with replacement from the pooled sample $\{\mathcal{S}_0, \mathcal{S}_1\}$ with the probability of sampling a specific x_i being

$$\Pr(x_i \in \mathcal{BS}_1) = \begin{cases} \dfrac{\hat{f}_{X|T}(x_i, T = 1)}{\hat{f}_{X|T}(x_i, T = 0)} \cdot \dfrac{1}{\sum_{i \in \mathcal{S}_0} \frac{\hat{f}_{X|T}(x_i,T=1)}{\hat{f}_{X|T}(x_i,T=0)}} \cdot \dfrac{N_0}{N_0 + N_1} & \text{if } i \in \mathcal{S}_0 \\ \dfrac{1}{N_0 + N_1} & \text{if } i \in \mathcal{S}_1 \end{cases} \tag{6}$$

   (c) Calculate and store the test statistic (3) using $\mathcal{BS}_0$ and $\mathcal{BS}_1$ instead of $\mathcal{S}_0$ and $\mathcal{S}_1$.

4. The p-value is the percentile rank of ck in the bootstrap distribution from step 3.