The conditional permutation test

Thomas B. Berrett∗, Yi Wang†, Rina Foygel Barber†, Richard J. Samworth∗

July 17, 2018

Abstract

We propose a general new method, the conditional permutation test, for testing the conditional independence of variables $X$ and $Y$ given a potentially high-dimensional random vector $Z$ that may contain confounding factors. The proposed test permutes entries of $X$ non-uniformly, so as to respect the existing dependence between $X$ and $Z$ and thus account for the presence of these confounders. Like the conditional randomization test of Candès et al. [5], our test relies on the availability of an approximation to the distribution of $X \mid Z$—while Candès et al. [5]'s test uses this estimate to draw new $X$ values, for our test we use this approximation to design an appropriate non-uniform distribution on permutations of the $X$ values already seen in the true data. We provide an efficient Markov Chain Monte Carlo sampler for the implementation of our method, and establish bounds on the Type I error in terms of the error in the approximation of the conditional distribution of $X \mid Z$, finding that, for the worst case test statistic, the inflation in Type I error of the conditional permutation test is no larger than that of the conditional randomization test. We validate these theoretical results with experiments on simulated data and on the Capital Bikeshare data set.

1 Introduction

Independence is a central notion in statistical model building, as well as being a foundational concept for much of statistical theory. Originating with Francis Galton’s work on correlation at the end of the 19th century [19], many measures of dependence have been proposed, including mutual information, the Hilbert–Schmidt independence criterion, and distance covariance [6, 10, 25]; see also [11] for an overview. Simultaneously, a great deal of research effort has gone into developing several different tests...
of independence, for example based on ranks, kernel methods, copulas, and nearest neighbours [26, 13, 12, 4]. Permutation tests are particularly attractive due to their simplicity and their ability to control the Type I error (i.e. the false positive rate) without any distributional assumptions.

In practice, it is often conditional independence that is in fact of primary interest [7]. For instance, in generalized linear models for a response $Y \in \mathbb{R}$ regressed on a high-dimensional feature vector $(X, Z) = (X, Z_1, \ldots, Z^p) \in \mathbb{R}^{p+1}$, the regression coefficient on feature $X$ is zero if and only if $Y$ and $X$ are conditionally independent given the remaining $p$ features, $Z = (Z_1, \ldots, Z^p)$. In this paper, we will study the general problem of testing $X \perp \perp Y \mid Z$. We are typically interested in the setting where $X$ and $Y$ are one-dimensional while $Z$ is a high-dimensional set of confounding variables that we would like to control for, but our results are not specific to this setting.

Within standard parametric regression models, conditional independence tests are well-developed; unfortunately, however, they fail to control Type I error under model misspecification. In fact, the very recent work of Shah and Peters [17] has shown that, without placing some assumptions on the joint distribution of $(X, Y, Z)$, conditional testing is effectively impossible—when $(X, Y, Z)$ is continuously distributed, they prove that there is no conditional independence test that both (1) controls Type I error over any null distribution (i.e. any distribution of $(X, Y, Z)$ with $X \perp \perp Y \mid Z$), and (2) has better than random power against even one alternative hypothesis.

Our work seeks to complement this fundamental result of Shah and Peters [17] by demonstrating that, given some additional knowledge, namely an approximation to the conditional distribution of $X$ given $Z$, one can in fact derive conditional independence tests that are approximately valid in finite samples, and that have non-trivial power.

1.1 Summary of contributions

In this paper, we introduce a new method, called the conditional permutation test (CPT), which is inspired by the conditional randomization test (CRT) of Candès et al. [5]. The CPT modifies the standard permutation test by using available distributional information to account correctly for the confounding variables $Z$, which leads to a non-uniform distribution over the set of possible permutations $\pi$ on the $n$ observations in our data set, and restores Type I error control.

Implementing the CPT is a challenging problem since we are sampling from a highly non-uniform distribution over the space of $n!$ permutations, but we propose a Monte Carlo sampler that yields an efficient implementation of the test. We additionally develop theoretical results examining the robustness of both the CPT and the CRT to slight errors in modeling assumptions, proving that Type I error is only slightly inflated

---

1In the regression literature, it is more common to use the notation of regressing $Y$ on $(X^1, \ldots, X^p)$, and testing whether the coefficient on feature $X^j$ is zero after controlling for the remaining features $X^{-j} = (X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p)$; this $X^j$ and $X^{-j}$ correspond to our $X$ and $Z$, respectively.
in both tests when our available distributional information is only approximately correct. In fact, in the worst case, Type I error is always less inflated for the new CPT method as compared to the CRT. Our empirical results verify the greater robustness of the CPT, while maintaining comparable power in a range of scenarios.

2 Background

In this section, we briefly summarize several existing approaches to the problem of testing for dependence between $X$ and $Y$ in the presence of confounding variables. Before beginning, it will be helpful to define some brief notation. Throughout, we will assume that the data consists of i.i.d. data points $(X_i, Y_i, Z_i) \in X \times Y \times Z$ for $i = 1, \ldots, n$, and will write $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, and $Z = (Z_1, \ldots, Z_n)$.

2.1 Permutation tests

One key reason why handling conditional independence in nonparametric contexts is so challenging, is that the permutation approaches that are so effective for testing unconditional independence, $X \perp Y$, cannot be directly applied when we seek to test conditional independence, $X \perp Y \mid Z$. This is because it may be the case that the null hypothesis $H_0 : X \perp Y \mid Z$ is true, but $X$ and $Y$ are highly marginally dependent due to correlation induced via each variable’s dependence on $Z$. Under this null, if we sample a permutation $\pi$ of $\{1, \ldots, n\}$ uniformly at random, then the permuted data set $(X_{\pi(1)}, Y_1, \ldots, (X_{\pi(n)}, Y_n)$ may have a very different distribution from the original data set $(X_1, Y_1, \ldots, (X_n, Y_n)$, due to the confounding effect of $Z$.

In certain settings, in particular where $Z$ is categorical, there is a simple and well-known fix for this problem: we can group the observations according to their value of $Z$, and then permute within groups. For example, if $Z \in \{0, 1\}$ is binary, we could draw a permutation $\pi$ that permutes the $X_i$’s within the set of indices $\{i : Z_i = 0\}$, and separately permutes the $X_i$’s within the set $\{i : Z_i = 1\}$. However, this strategy cannot be applied directly in the case where $Z$ is continuously distributed, or where $Z$ is discrete but with few repeated values (note that when $Z$ is high-dimensional, even if it is discrete, each observation $i$ will typically have a unique feature vector $Z_i$). In these settings, it is common to use a binning strategy, where first $Z$ is discretized to fall into finitely many bins, and then the “permute within groups” strategy is deployed. However, Type I error control is no longer guaranteed, since the null hypothesis $H_0 : X \perp Y \mid Z$ does not imply that $X \perp Y \mid (Z \in \text{bin } b)$; the best we can usually hope for is that the latter statement would be approximately true under the null. Furthermore, in a high-dimensional setting, choosing these bins can itself be very challenging.
2.2 The conditional randomization test

The conditional randomization test (CRT), proposed by Candès et al. [5], works in a setting where no assumptions are made about the distribution of the response variable $Y$, but instead, it is assumed that the conditional distribution of $X$ given $Z$ is known. In practice, in semi-supervised learning settings where unlabeled data $(X, Z)$ are easier to obtain than labeled data $(X, Y, Z)$, it may be possible to obtain a very accurate estimate of the conditional distribution $X \mid Z$, but testing for independence with $Y$ remains challenging due to limited sample size of the labeled data.

The CRT works by sampling a new copy of the $X$ values in the data set. Letting $Q(\cdot \mid z)$ denote the distribution of $X$ given $Z = z$, conditional on $Z_1, \ldots, Z_n$, the CRT draws

$$X_i^{(1)} \sim Q(\cdot \mid Z_i),$$

independently for each $i = 1, \ldots, n$, and independently of the observed $X_i$’s and $Y_i$’s. (In the special case where $X$ is binary, earlier work by Rosenbaum [14] proposed a related test, referred to as a “conditional permutation test” but which in fact resamples $X$ by estimating $P \{X = 1 \mid Z\}$ with a logistic model.)

Under the null hypothesis $H_0$ that $X \perp \perp Y \mid Z$, we see that

$$(X \mid Y = y, Z = z) \overset{d}{=} (X \mid Z = z) \sim Q(\cdot \mid z),$$

where $\overset{d}{=}$ denotes equality in distribution. This means that

$$(X^{(1)}, Y, Z) \overset{d}{=} (X, Y, Z) \text{ under } H_0,$$

where $X^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)})$. Any large differences between these two triples—for instance, if $Y$ is highly correlated with $X$ but not with $X^{(1)}$—can therefore be interpreted as evidence against the null hypothesis. In order to construct a test of $H_0$, then, the CRT repeats this process $M$ times, sampling

$$(X_i^{(m)} \mid X, Y, Z) \sim Q(\cdot \mid Z_i),$$

independently for $i = 1, \ldots, n$ and $m = 1, \ldots, M$ to form control vectors $X^{(1)}, \ldots, X^{(M)}$. Under the null hypothesis, the triples $(X, Y, Z)$, $(X^{(1)}, Y, Z), \ldots, (X^{(M)}, Y, Z)$ are all identically distributed; in fact, they are exchangeable. For any statistic $T = T(X, Y, Z)$ that is chosen in advance (or, at least, without looking at $X$), the values

$$T(X, Y, Z), T(X^{(1)}, Y, Z), \ldots, T(X^{(M)}, Y, Z)$$

are therefore exchangeable as well. We can compute a $p$-value by ranking the value obtained from the true $X$ vector against the values obtained from the CRT’s copies:

$$p = \frac{1 + \sum_{m=1}^{M} 1 \{T(X^{(m)}, Y, Z) \geq T(X, Y, Z)\}}{1 + M}. $$
The exchangeability of the random variables in (1) ensures that this is a valid p-value, i.e. it satisfies \( P \{ p \leq \alpha \} \leq \alpha \) for all \( \alpha \in [0, 1] \).

The “model-X knockoffs” framework of Candès et al. [5] also extends the CRT technique to the high-dimensional variable selection setting, where each of \( p \) features is tested in turn for conditional independence with the response \( Y \), with the goal of false discovery rate control. In this framework, only a single copy of each feature is created. The robustness of the model-X knockoffs method, with respect to errors in the conditional distributions used to construct the knockoff copies of each feature (analogous to the \( X^{(m)} \)’s above), was studied by Barber et al. [1].

### 2.3 Other tests of conditional independence

Before introducing our new work, we give a brief overview of some additional conditional independence testing methods proposed in the literature. Many methods assume some parametric model for the response \( Y \), such as a linear model, \( Y = \alpha X + \beta^\top Z + \text{(noise)} \), in which case the problem reduces to testing whether \( \alpha = 0 \). This can be tested by, for instance, computing an estimate \( \hat{\beta} \) and testing whether the residual \( Y - \hat{\beta}^\top Z \) is correlated with \( X \). Belloni et al. [2] propose a variant on this approach, which assumes approximate linear models for both \( Y \) and \( X \). Their method regresses both \( X \) and \( Y \) on \( Z \), then tests for correlation between the two resulting residual vectors; this “double regression” offers superior performance by removing much of the bias coming from errors in estimating the effect of \( Z \). Shah and Peters [17] consider a more general double regression framework, assuming that the conditional means \( \mathbb{E}[X | Z = z] \) and \( \mathbb{E}[Y | Z = z] \) can be estimated at a sufficiently fast rate.

Away from the regression setting, many proposed methods are based on using kernel representations or low-dimensional projections of the data. Tests based on embedding the data into reproducing kernel Hilbert spaces are studied in, for example, Fukumizu et al. [9], Zhang et al. [27] and Strobl et al. [20]. Other works use permutations of the data, including Doran et al. [8] and Sen et al. [16], where the methods have the flavor of binning \( Z \) and then permuting within groups. The use of partial copulas was studied by Bergsma [3] and Song [18]. There is also a large literature on extending measures of marginal independence to the conditional setting, including partial distance covariance [24]; conditional mutual information [15]; characteristic functions [21]; Hellinger distances [22]; and smoothed empirical likelihoods [23].

### 3 The conditional permutation test (CPT)

Recall that the conditional randomization test (CRT) [5] creates copies \( X^{(m)} \) of the vector \( X \) sampled under the null hypothesis that \( X \perp \perp Y | Z \), by drawing

\[
X^{(m)} | X, Y, Z \sim Q^m(\cdot | Z), \quad \text{independently for } m = 1, \ldots, M, \tag{2}
\]
where we define $Q^n(\cdot|Z) := Q(\cdot|Z_1) \times \cdots \times Q(\cdot|Z_n)$. This mechanism creates copies $X^{(1)}, \ldots, X^{(M)}$ that are exchangeable with the original vector $X$ under the null hypothesis that $X \perp \perp Y \mid Z$.

Our proposed method, the conditional permutation test (CPT), is a variant on the CRT, with $X^{(1)}, \ldots, X^{(M)}$ drawn as in (2) but under the constraint that each $X^{(m)}$ must be a permutation of the original vector $X$.

In order to make this concrete, we first need to define some notation. It is convenient to place a total ordering on $X$; this can be arbitrary and does not affect computations, but in the case $X = \mathbb{R}$ the usual ordering is obviously most natural. It then makes sense to let $X^{(1)} \leq \cdots \leq X^{(n)}$ denote the order statistics of $X$ and write $X^{()}$ for the vector of order statistics. Next, for any permutation $\pi$ of the indices $[n] := \{1, \ldots, n\}$, we define the vector $X^{(\pi)} := (X^{(\pi(1))}, \ldots, X^{(\pi(n))})$.

In other words, $X^{(\pi)}$ is the vector whose order statistics are given by $X^{()}$, and whose ranks (in ascending order) are determined by the permutation $\pi$.

We are now ready to define the CPT. After observing $X, Y, Z$, we use the following discrete distribution to draw $X^{(m)}$:

$$
P \{ X^{(m)} = X^{(\pi)} \mid X^{()}, Y, Z \} \propto q^n(X^{(\pi)} \mid Z).$$

Here we let $q(\cdot|z)$ be the density of the distribution $Q(\cdot|z)$ (i.e. $q(\cdot|z)$ is the conditional density of $X$ given $Z = z$), with respect to some base measure $\nu$ on $X$ that does not depend on $z$. We write $q^n(\cdot|Z) := q(\cdot|Z_1) \cdots q(\cdot|Z_n)$ to denote the product density. Note that we are not assuming a continuous distribution necessarily; the base measure may be discrete, allowing $X$ to be discrete as well.

As before, the copies $X^{(1)}, \ldots, X^{(M)}$ are drawn independently from (3), conditional on $X^{()}, Y, Z$. These control vectors $X^{(m)}$ are then used exactly as for the CRT—given some predefined statistic $T = T(X, Y, Z)$, our p-value is given by

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1} \{ T(X^{(m)}, Y, Z) \geq T(X, Y, Z) \}}{1 + M}.$$  

Comparing the mechanisms (2) for the CRT and (3) for the CPT, we see that in both cases the copies $X^{(m)}$ are drawn according to the conditional distribution $Q^n(\cdot|Z)$, but for the CPT we restrict to vectors $X^{(m)}$ that are permutations of the original $X$ vector.

The following theorem verifies that this procedure yields a valid test of $H_0$. (The proof of this theorem, and all other proofs, are given in Appendix A.)

**Theorem 1.** Assume that $H_0 : X \perp \perp Y \mid Z$ is true, and that the conditional distribution of $X \mid Z$ is given by $Q(\cdot|Z)$. Then under the CPT, the $M + 1$ triples

$$(X, Y, Z), (X^{(1)}, Y, Z), \ldots, (X^{(M)}, Y, Z)$$

where we define $Q^n(\cdot|Z) := Q(\cdot|Z_1) \times \cdots \times Q(\cdot|Z_n)$. This mechanism creates copies $X^{(1)}, \ldots, X^{(M)}$ that are exchangeable with the original vector $X$ under the null hypothesis that $X \perp \perp Y \mid Z$. Our proposed method, the conditional permutation test (CPT), is a variant on the CRT, with $X^{(1)}, \ldots, X^{(M)}$ drawn as in (2) but under the constraint that each $X^{(m)}$ must be a permutation of the original vector $X$.

In order to make this concrete, we first need to define some notation. It is convenient to place a total ordering on $X$; this can be arbitrary and does not affect computations, but in the case $X = \mathbb{R}$ the usual ordering is obviously most natural. It then makes sense to let $X^{(1)} \leq \cdots \leq X^{(n)}$ denote the order statistics of $X$ and write $X^{()}$ for the vector of order statistics. Next, for any permutation $\pi$ of the indices $[n] := \{1, \ldots, n\}$, we define the vector $X^{(\pi)} := (X^{(\pi(1))}, \ldots, X^{(\pi(n))})$.

In other words, $X^{(\pi)}$ is the vector whose order statistics are given by $X^{()}$, and whose ranks (in ascending order) are determined by the permutation $\pi$.

We are now ready to define the CPT. After observing $X, Y, Z$, we use the following discrete distribution to draw $X^{(m)}$:

$$P \{ X^{(m)} = X^{(\pi)} \mid X^{()}, Y, Z \} \propto q^n(X^{(\pi)} \mid Z).$$

Here we let $q(\cdot|z)$ be the density of the distribution $Q(\cdot|z)$ (i.e. $q(\cdot|z)$ is the conditional density of $X$ given $Z = z$), with respect to some base measure $\nu$ on $X$ that does not depend on $z$. We write $q^n(\cdot|Z) := q(\cdot|Z_1) \cdots q(\cdot|Z_n)$ to denote the product density. Note that we are not assuming a continuous distribution necessarily; the base measure may be discrete, allowing $X$ to be discrete as well.

As before, the copies $X^{(1)}, \ldots, X^{(M)}$ are drawn independently from (3), conditional on $X^{()}, Y, Z$. These control vectors $X^{(m)}$ are then used exactly as for the CRT—given some predefined statistic $T = T(X, Y, Z)$, our p-value is given by

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1} \{ T(X^{(m)}, Y, Z) \geq T(X, Y, Z) \}}{1 + M}.\quad (4)$$

Comparing the mechanisms (2) for the CRT and (3) for the CPT, we see that in both cases the copies $X^{(m)}$ are drawn according to the conditional distribution $Q^n(\cdot|Z)$, but for the CPT we restrict to vectors $X^{(m)}$ that are permutations of the original $X$ vector.

The following theorem verifies that this procedure yields a valid test of $H_0$. (The proof of this theorem, and all other proofs, are given in Appendix A.)

**Theorem 1.** Assume that $H_0 : X \perp \perp Y \mid Z$ is true, and that the conditional distribution of $X \mid Z$ is given by $Q(\cdot|Z)$. Then under the CPT, the $M + 1$ triples

$$(X, Y, Z), (X^{(1)}, Y, Z), \ldots, (X^{(M)}, Y, Z)$$

where we define $Q^n(\cdot|Z) := Q(\cdot|Z_1) \times \cdots \times Q(\cdot|Z_n)$. This mechanism creates copies $X^{(1)}, \ldots, X^{(M)}$ that are exchangeable with the original vector $X$ under the null hypothesis that $X \perp \perp Y \mid Z$. Our proposed method, the conditional permutation test (CPT), is a variant on the CRT, with $X^{(1)}, \ldots, X^{(M)}$ drawn as in (2) but under the constraint that each $X^{(m)}$ must be a permutation of the original vector $X$.

In order to make this concrete, we first need to define some notation. It is convenient to place a total ordering on $X$; this can be arbitrary and does not affect computations, but in the case $X = \mathbb{R}$ the usual ordering is obviously most natural. It then makes sense to let $X^{(1)} \leq \cdots \leq X^{(n)}$ denote the order statistics of $X$ and write $X^{()}$ for the vector of order statistics. Next, for any permutation $\pi$ of the indices $[n] := \{1, \ldots, n\}$, we define the vector $X^{(\pi)} := (X^{(\pi(1))}, \ldots, X^{(\pi(n))})$.

In other words, $X^{(\pi)}$ is the vector whose order statistics are given by $X^{()}$, and whose ranks (in ascending order) are determined by the permutation $\pi$.

We are now ready to define the CPT. After observing $X, Y, Z$, we use the following discrete distribution to draw $X^{(m)}$:

$$P \{ X^{(m)} = X^{(\pi)} \mid X^{()}, Y, Z \} \propto q^n(X^{(\pi)} \mid Z).$$

Here we let $q(\cdot|z)$ be the density of the distribution $Q(\cdot|z)$ (i.e. $q(\cdot|z)$ is the conditional density of $X$ given $Z = z$), with respect to some base measure $\nu$ on $X$ that does not depend on $z$. We write $q^n(\cdot|Z) := q(\cdot|Z_1) \cdots q(\cdot|Z_n)$ to denote the product density. Note that we are not assuming a continuous distribution necessarily; the base measure may be discrete, allowing $X$ to be discrete as well.

As before, the copies $X^{(1)}, \ldots, X^{(M)}$ are drawn independently from (3), conditional on $X^{()}, Y, Z$. These control vectors $X^{(m)}$ are then used exactly as for the CRT—given some predefined statistic $T = T(X, Y, Z)$, our p-value is given by

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1} \{ T(X^{(m)}, Y, Z) \geq T(X, Y, Z) \}}{1 + M}.\quad (4)$$

Comparing the mechanisms (2) for the CRT and (3) for the CPT, we see that in both cases the copies $X^{(m)}$ are drawn according to the conditional distribution $Q^n(\cdot|Z)$, but for the CPT we restrict to vectors $X^{(m)}$ that are permutations of the original $X$ vector.

The following theorem verifies that this procedure yields a valid test of $H_0$. (The proof of this theorem, and all other proofs, are given in Appendix A.)

**Theorem 1.** Assume that $H_0 : X \perp \perp Y \mid Z$ is true, and that the conditional distribution of $X \mid Z$ is given by $Q(\cdot|Z)$. Then under the CPT, the $M + 1$ triples

$$(X, Y, Z), (X^{(1)}, Y, Z), \ldots, (X^{(M)}, Y, Z)$$

where we define $Q^n(\cdot|Z) := Q(\cdot|Z_1) \times \cdots \times Q(\cdot|Z_n)$.
are exchangeable. In particular, this implies that for any statistic \( T : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R} \), the p-value defined in (4) is valid, satisfying \( \mathbb{P}\{p \leq \alpha\} \leq \alpha \) for any desired Type I error rate \( \alpha \in [0, 1] \).

In other words, this theorem states that if the conditional distribution \( Q(\cdot | z) \) is known, then the CPT is a valid test of conditional independence. The analogous result for the CRT is proved in Candès et al. [5, Lemma 4.1].

### 3.1 Comparing CPT and CRT

**Use of marginal distribution of X** Both the CPT and the CRT provide Type I error control at the desired level \( \alpha \), as long as \( Q(\cdot | Z) \) is indeed the exact conditional distribution of \( X | Z \), rather than an estimate. But how will the two tests compare in their control of Type I error when \( Q(\cdot | Z) \) is not the correct conditional distribution or their power to detect when \( H_0 \) is false? In terms of how the tests are run, the difference can be described as follows: while both tests use the (true or estimated) conditional distribution \( Q(\cdot | Z) \), the CPT additionally uses the marginal distribution of the observed data vector \( X \). Intuitively, using this additional information can in some cases make the copies \( X^{(m)} \) more similar to the original \( X \), than for the CRT. Therefore, the CPT may be somewhat less likely to reject \( H_0 \), which could lead to lower Type I error if \( H_0 \) is true, or reduced power to detect when \( H_0 \) is false. In Section 5, we will develop theory to examine the two tests’ robustness to errors in estimating the conditional distribution \( Q(\cdot | Z) \), and we will compare the tests in terms of both Type I error and power in experiments in Section 6.

**Invariance to base measure** Since the CPT works only over permutations of the same set of \( X \) values, it follows that it is invariant to changes in the base measure on \( \mathcal{X} \). To make this concrete, suppose that \( q_1(\cdot | z) \) is another conditional density, with the property that there exist functions \( h(\cdot) \), \( c(\cdot) \) such that \( q_1(x | z) = q(x | z)h(x)c(z) \) for all \( x \in \mathcal{X} \) and all \( z \in \mathcal{Z} \). (Here we can think of \( h(x) \) as changing the base measure on \( \mathcal{X} \), while \( c(z) \) adjusts the normalizing constants as needed.)

If this is the case, then running the CPT with \( q_1 \) in place of \( q \) will have no effect on the outcome—this is because we can calculate

\[
q_1^n(X(\pi)|Z) = \prod_{i=1}^n q(X(\pi(i)) | Z_i)h(X(\pi(i)))c(Z_i) = q^n(X(\pi)|Z) \cdot \prod_{i=1}^n h(X_i)c(Z_i).
\]

The first term, \( q^n(X(\pi)|Z) \), is the same as for the CPT run with conditional density \( q \), while the second term, \( \prod_{i=1}^n h(X_i)c(Z_i) \), does not depend on the permutation \( \pi \) and therefore does not affect the resulting distribution of the sampled permutations.

This means that the CPT is a valid test, i.e. the result of Theorem 1 holds, even if the conditional density \( q(\cdot | z) \) is correct only up to a change in base measure. In other
words, Theorem \[\square\] holds whenever the conditional distribution \( Q(\cdot | Z) \) has a density of the form \( q(x | z) h(x) c(z) \), for some functions \( h(\cdot), c(\cdot) \). Indeed, in some settings, it may be substantially simpler to estimate the conditional density only up to base measure—for instance, we can consider a semiparametric model with a conditional density of the form \( \exp\{x \cdot z^\top \theta - f(x) - g(z)\} \), in which case the CPT would only need to estimate the parametric component \( \theta \). In contrast, running the CRT requires being able to sample from the conditional distribution \( Q(\cdot | Z) \), so we would need to approximate the full conditional density.

4 Sampling algorithms for the CPT

In order to run the CPT, we need to be able to sample permutations from the distribution given in \(\square\). We now turn to the problem of generating such samples efficiently.

One simple approach would be to run a Metropolis–Hastings algorithm with a proposal distribution that, from a current state \( \pi \), draws its proposed permutation \( \pi' \) uniformly at random. For even a moderate \( n \), however, the acceptance odds ratio

\[
\frac{\prod_{i=1}^{n} q(X_{\pi'(i)} | Z_i)}{\prod_{i=1}^{n} q(X_{\pi(i)} | Z_i)}
\]

will be extremely low for nearly all permutations \( \pi' \) (unless, of course, the dependence of \( X \) on \( Z \) is very weak). In other words, a uniformly drawn permutation \( \pi' \) is not likely to lead to a plausible vector of \( X \) values, leading to slow mixing times.

As a second attempt, we can consider a different proposal distribution: from the current state \( \pi \), we propose the permutation \( \pi' = \pi \circ \sigma_{ij} \), where \( \sigma_{ij} \) is the permutation that swaps indices \( i \) and \( j \), which are drawn at random. The acceptance odds ratio \(\square\) now simplifies to

\[
\frac{q(X_{\pi(j)} | Z_i) \cdot q(X_{\pi(i)} | Z_j)}{q(X_{\pi(i)} | Z_i) \cdot q(X_{\pi(j)} | Z_j)}.
\]

The probability of accepting a swap will now be reasonably high; however, each step can only alter two of the \( n \) indices, again leading to slow mixing times.

4.1 A parallelized pairwise sampler

To address these issues, we propose a parallelized version of this pairwise algorithm. At each step, we first draw \([n/2]\) disjoint pairs of indices from \([n]\). Next, independently and in parallel for each pair, we decide whether or not to swap this pair \((i, j)\), according to the odds ratio \(\square\). This sampler is defined formally in Algorithm \[\square\].

The next theorem verifies that the resulting Markov chain yields the desired stationary distribution:
Algorithm 1 Parallelized pairwise sampler for the CPT

**Input:** Initial permutation $\Pi^{[0]}$.

for $t = 1, 2, \ldots$ do

Sample uniformly without replacement from $[n]$ to obtain disjoint pairs $(i_{t,1}, j_{t,1}), \ldots, (i_{t,\lfloor n/2 \rfloor}, j_{t,\lfloor n/2 \rfloor})$.

Draw independent Bernoulli variables $B_{t,1}, \ldots, B_{t,\lfloor n/2 \rfloor}$ with odds ratios

$$\frac{\Pr\{B_{t,k} = 1\}}{\Pr\{B_{t,k} = 0\}} = \frac{q(X_{\Pi^{[t-1]}(j_{t,k})} | Z_{i_{t,k}}) \cdot q(X_{\Pi^{[t-1]}(i_{t,k})} | Z_{j_{t,k}})}{q(X_{\Pi^{[t-1]}(i_{t,k})} | Z_{i_{t,k}}) \cdot q(X_{\Pi^{[t-1]}(j_{t,k})} | Z_{j_{t,k}})}.$$  \hspace{1cm} (7)

Define $\Pi^{[t]}$ by swapping $\Pi^{[t-1]}(i_{t,k})$ and $\Pi^{[t-1]}(j_{t,k})$ for each $k$ with $B_{t,k} = 1$.

end for

Theorem 2. For every initial permutation $\Pi^{[0]}$, the stationary distribution of the Markov chain defined in Algorithm 1 is unique, and is equal to the distribution $\Pr\{\Pi|X\} \cdot \Pr\{X\}$ of the ranks $\Pi$ conditional on $X$, $Y$, $Z$.

In particular, if $\Pi$ denotes the vector of ranks of $X$ and Algorithm 1 is initialized at $\Pi$, then under the null hypothesis that $X \perp Y | Z$, the initial value is already drawn from the stationary distribution; therefore the correct distribution is preserved at each step of the sampler.

In practice, we want to draw $M$ copies, $X^{(m)}$ for $m = 1, \ldots, M$, and we need to ensure that the original data $X$ and each of the $M$ permutations $X^{(m)}$ are all exchangeable with each other. If we sample the permuted vectors $X^{(1)}, \ldots, X^{(M)}$ sequentially, by running Algorithm 1 for $S \cdot M$ steps and extracting one copy $X^{(m)}$ after each round of $S$ steps, then we would not achieve exchangeability, since there would be some correlation between adjacent copies in this sequence. (Of course, in practice, if the number of steps $S$ is chosen to be large, then the violation of exchangeability would be very mild.)

Instead, we can construct an exchangeable sampling mechanism with the following algorithm:

Algorithm 2 Exchangeable sampler for multiple draws from the CPT

**Input:** Initial permutation $\Pi_{\text{init}}$ and integer $S \geq 1$.

Define $\Pi^{[t]}$ by running Algorithm 1 initialized at $\Pi^{[0]} = \Pi_{\text{init}}$ for $S$ steps.

for $m = 1, \ldots, M$ (independently for each $m$) do

Define $\Pi^{(m)}$ by running Algorithm 1 initialized at $\Pi^{[0]} = \Pi^{[t]}$ for $S$ steps.

end for
Algorithm 2 provides an exchangeable sampling mechanism, since the permutation $\Pi$ is at the “center”, lying $S$ steps away from each of the permutations $\Pi, \Pi^{(1)}, \ldots, \Pi^{(M)}$. The following result verifies exchangeability:

**Theorem 3.** Let $X_{(i)}$ and $\Pi$ be the order statistics and ranks of $X$, as defined previously, so that $X = X_{(\Pi)}$. Let $\Pi^{(1)}, \ldots, \Pi^{(M)}$ be the output of Algorithm 2 and let $X^{(m)} = X_{(\Pi^{(m)})}$ for each $m = 1, \ldots, M$. Assume that the null hypothesis that $X \independent Y \mid Z$ holds, and the conditional distribution of $X \mid Z$ is given by $Q(\cdot \mid Z)$, so that the distribution of $\Pi$ conditional on $X_{(i)}, Y, Z$ is given by $Q$. Then the triples $(X, Y, Z), (X^{(1)}, Y, Z), \ldots, (X^{(M)}, Y, Z)$ are exchangeable.

This result ensures that the results of Theorem 1 hold when the permuted vectors $X^{(1)}, \ldots, X^{(M)}$ are obtained via the exchangeable sampler.

### 5 Robustness of the CPT and CRT

We next consider whether the CPT and CRT, based on resampling $X$ from a known or estimated conditional distribution given $Z$, are robust to slight errors in this distribution. Suppose that the conditional distribution $Q(\cdot \mid Z)$ that we use for sampling when running the CPT or CRT is only an approximation to the true conditional, denoted by $Q_\ast(\cdot \mid Z)$. In this section we provide bounds on the excess Type I error of the CPT and CRT as a function of the difference between the true conditional $Q_\ast$ and its approximation $Q$. Throughout, we will assume that the statistic $T : X^n \times Y^n \times Z^n \to \mathbb{R}$ used in the test, as well as the approximation $Q$ to the conditional distribution, are chosen independently of $X, Y$. For instance, in many applications, we may have access to unlabeled data, i.e. draws of $(X, Z)$ without $Y$, which we can use to construct an estimate $Q$.

Our first result demonstrates that, conditional on $Y, Z$, the excess Type I error of both the CPT and the CRT is bounded by the total variation distance between $Q_\ast$ and $Q$. (For any two distributions $Q_1, Q_2$ defined on the same probability space, the total variation distance is defined as $d_{TV}(Q_1, Q_2) = \sup_A |Q_1(A) - Q_2(A)|$, where the supremum is taken over all measurable sets.)

**Theorem 4.** Assume that $H_0 : X \independent Y \mid Z$ is true, and that the conditional distribution of $X \mid Z$ is given by $Q_\ast(\cdot \mid Z)$. For a fixed integer $M \geq 1$, let $X^{(1)}, \ldots, X^{(M)}$ be copies of $X$ generated from the CRT (2) or from the CPT (3) using an estimate $Q$ of the true conditional distribution $Q_\ast$.

Then, for any desired Type I error rate $\alpha \in [0, 1]$,

$$\mathbb{P}\{p \leq \alpha \mid Y, Z\} \leq \alpha + d_{TV}(Q^n_\ast(\cdot \mid Z), Q^n(\cdot \mid Z)), $$

where $p$ is the $p$-value computed in (4), and the probability is taken with respect to the distribution of $X, X^{(1)}, \ldots, X^{(M)}$ conditional on $Y, Z$. 


Of course, we can also bound the Type I error rate unconditionally, with
\[ P\{p \leq \alpha\} \leq \alpha + \mathbb{E}\left[d_{TV}(Q^*_n(\cdot|Z), Q^n(\cdot|Z))\right], \]
which we obtain from the result above by marginalizing over \(Y, Z\).

This result ensures that, if \(Q\) is a good approximation to \(Q_*\), then both the CPT and CRT will have at most a mild increase in their Type I error. Of course, Theorem 4 is a worst-case result, proved with respect to an arbitrary statistic \(T\) which may be chosen adversarially so as to be maximally sensitive to errors in estimating the true conditional distribution \(Q_*\). In practice, we might expect that the simple statistics \(T\) that we would most often use, such as correlation between \(X\) and \(Y\), could be more robust to errors than the theorem suggests.

While Theorem 4 provides an upper bound on the Type I error for both the CPT and the CRT, we do not yet have a comparison between the two. The following theorem proves that, for the case of the CRT, the upper bound is in fact tight when the number of copies \(X^{(1)}, \ldots, X^{(M)}\) is large:

**Theorem 5.** Under the setting and assumptions of Theorem 4, there exists a statistic \(T : X^n \times Y^n \times Z^n \to \mathbb{R}\) such that, for the CRT\(^2\)
\[
\sup_{\alpha \in [0, 1]} \left( P\{p \leq \alpha \mid Y, Z\} - \alpha \right) \geq d_{TV}(Q^*_n(\cdot|Z), Q^n(\cdot|Z)) - 0.5(1 + o(1))\sqrt{\frac{\log(M)}{M}}
\]
as \(M \to \infty\).

In other words, if we use the statistic \(T\) that is best able to detect errors in our conditional distribution, then the excess Type I error of the CRT is exactly equal to \(d_{TV}(Q^*_n(\cdot|Z), Q^n(\cdot|Z))\) (up to a vanishing factor), and therefore is at least as high as that of the CPT under any statistic. While this lower bound applies only to a specific \(T\), and does not guarantee that the excess error of the CRT will bound that of the CPT when both tests use some other statistic \(T\), in Section 6 we will see that, empirically, the CPT often yields a far lower Type I error than the CRT in simulations.

To see one example of a setting where Theorem 4 can be applied, suppose that the conditional density estimate \(q\) is uniformly accurate, in the sense that there exists \(\delta > 0\) such that
\[
\left| \log \frac{q(x|z)}{q_*(x|z)} \right| \leq \delta n^{-1/2} \quad \text{for all } x \in X \text{ and all } z \in Z,
\]
where \(q_*(\cdot|z)\) denotes the density of \(Q(\cdot|z)\) with respect to \(\nu\). In this case, we have
\[
d_{TV}(Q^*_n(\cdot|Z), Q^n(\cdot|Z))^2 \leq \frac{1}{2} \mathbb{E}\left[ \log \frac{q^n(X | Z)}{q^n(X | Z)} \mid Z \right] = \frac{1}{2} \sum_{i=1}^n \mathbb{E}\left[ \log \frac{q_*(X_i|Z_i)}{q_*(X_i|Z_i)} \mid Z_i \right]
\]
\[ \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E}\left[ 1 - \frac{q(X_i|Z_i)}{q_*(X_i|Z_i)} + \frac{1 + o(1)}{2} \left( \delta n^{-1/2} \right)^2 \mid Z_i \right] = \frac{1 + o(1)}{4} \delta^2, \]

\(^2\)To be more precise with the constant, we can replace 0.5(1 + o(1)) with 2.5 for any \(M \geq 2\).
as \( n \to \infty \), where the first step holds by Pinsker’s inequality; the next-to-last step uses the Taylor expansion of \( \log(\cdot) \); and the last step holds since the \( i \)th expectation is taken by integrating against the density \( q_*(\cdot|Z_i) \). This calculation then bounds the worst-case inflation of Type I error according to Theorem 4. An important observation to make here is that, since the CPT is invariant to changes in the base measure on \( \mathcal{X} \), this bound on total variation holds even when the condition \( (8) \) is replaced with

\[
\left| h(x)c(z) \frac{q(x|z)}{q_*(x|z)} \right| \leq \delta n^{-1/2} \quad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z}.
\]

### 6 Empirical results

We next examine the empirical performance of the CPT and CRT on simulated data, and on real data from the Capital Bikeshare system. Code for reproducing all experiments is available on the authors’ websites.

#### 6.1 Simulated data

The results of Section 5 show that the CPT is more robust than the CRT to errors in the estimated conditional distribution \( Q(\cdot|Z) \), when the worst case test statistics \( T(X,Y,Z) \) are used. Our first aim here is to provide evidence to validate this result, and to show that this extra robustness is not only exhibited by the worst case test statistic but also for practical and simple choices of \( T \). Our second aim is to examine the power of the CPT and CRT to detect deviations from the null hypothesis.

In all of our simulations we set \( \alpha = 0.05 \) as the desired Type I error rate, and use marginal absolute correlation \( T(X,Y,Z) = |\text{Corr}(X,Y)| \) as our test statistic. We generate \( M = 500 \) copies of \( X \) under either CPT or CRT. To run the CPT, we use Algorithm 2 with \( S = 50 \) steps. All results are shown averaged over 1000 trials.

#### 6.1.1 Simulations under the null

First we test whether the CPT and CRT show large increases in Type I error when the conditional distribution estimate \( Q(\cdot|Z) \) is incorrect, in a setting where the null hypothesis \( H_0 : X \perp \! \! \! \perp Y \mid Z \) holds.

We will have \( X, Y \in \mathbb{R} \) and \( Z \in \mathbb{R}^p \) for \( p = 20 \). We first draw independent parameter vectors

\[
a, b \sim \mathcal{N}_p(0, I_p).
\]

The variables \( (X,Y,Z) \) are then generated as

\[
Z \sim \mathcal{N}_p(0, I_p), \quad X \mid Z \sim Q_*(\cdot|Z), \quad Y \mid X, Z \sim \mathcal{N}(p^{-1}a^TZ, 1),
\]

\[\text{Available at } \url{http://www.stat.uchicago.edu/~rina/cpt.html} \]

12
where \( Q_\star(\cdot|Z) \) will be specified below. (Note that \( Y \mid X, Z \) depends on \( Z \) only, since we are working under the null hypothesis that \( X \perp \perp Y \mid Z \).)

Throughout, the estimated conditional distribution of \( X \mid Z \) will be given by \( Q(\cdot|Z) = N(b^\top Z, 1) \), but this estimate might not be exactly correct. We will consider two common types of model misspecification: first, assuming a linear relationship between variables where this is in fact not the case, and second, assuming that the distribution of the noise is Gaussian with a common variance, when this is not a good approximation.

For the true conditional distribution \( Q_\star(\cdot|Z) \) of \( X \mid Z \), we consider:

1. Errors in the mean. \( Q_\star(\cdot|z) = N(\mu(z), 1) \), where \( \mu(\cdot) \) is given by:
   (a) Quadratic: \( \mu(z) = b^\top z + \theta(b^\top z)^2 \),
   (b) Cubic: \( \mu(z) = b^\top z - \theta(b^\top z)^3 \),
   (c) Tanh: \( \mu(z) = \tanh(\theta \cdot b^\top z)/\theta \).

2. Errors in the noise model. \( Q_\star(\cdot|z) = b^\top z + \text{(noise)} \), where the noise is drawn from one of the following distributions (normalized to have mean zero and variance 1):
   (a) Heavy-tailed: \( \text{(noise)} \sim t_{1/\theta}/\sqrt{\text{Var}(t_{1/\theta})} \), i.e. the \( t \) distribution with \( 1/\theta \) degrees of freedom,
   (b) Skewed: \( \text{(noise)} \sim \frac{\text{SN}(\theta)-\text{E}[\text{SN}(\theta)]}{\sqrt{\text{Var}(\text{SN}(\theta))}} \), where \( \text{SN}(\theta) \) is the skew-normal distribution with skew parameter \( \theta \), which has density \( 2\phi(t)\Phi(\theta t) \) at each \( t \in \mathbb{R} \),
   (c) Heteroskedastic: \( \text{(noise)} \sim N\left(0, \frac{|b^\top Z|^2}{\|b\|^2 \int_{-\infty}^{\infty} |t|^2 \phi(t) \, dt} \right) \). The normalization constant in the denominator means that the expected variance of the noise is 1 (with the expectation taken over \( Z \)).

In each case, \( \theta \geq 0 \) is the model misspecification parameter. Note that \( \theta = 0 \) corresponds to the case that \( Q(\cdot|Z) = Q_\star(\cdot|Z) \), i.e. the estimate is indeed correct, while larger values of \( \theta \) correspond to increasing errors.

Results The plots in Figures 1 and 2 show the results of these experiments when we have errors in the mean and in the noise model, respectively. As the null hypothesis, \( H_0 : X \perp \perp Y \mid Z \), is true in all of these experiments, we would hope for the probability of rejection to be close to the nominal level of \( \alpha = 0.05 \), at least when the model misspecification parameter \( \theta \) is not too large. For errors in the mean, in Figure 1 we see that in many cases the CPT is significantly more robust than the CRT. The \( \theta = 0 \) cases confirm that both tests achieve the nominal Type I error level \( \alpha = 0.05 \) when the assumed distribution \( Q \) is correct. For the simulations testing errors in the noise model, shown in Figure 2, we see that none of these experiments show any significant difference between the robustness of the CPT and the CRT, and both tests appear to have approximately the nominal size across a range of noise model error levels.
Figure 1: Simulation results for robustness to misspecification of the mean function. The figures show the probability of rejection (i.e. the Type I error rate), plotted against the model misspecification parameter $\theta$. The plots show the average rejection probability with standard error bars computed over 1000 trials for the CPT and CRT. The dashed line indicates the nominal level $\alpha = 0.05$.

Figure 2: Simulation results for robustness to misspecification of the noise model. Details as for Figure 1.
Figure 3: Simulation results testing power under the alternative. The figures show the probability of rejection (i.e. the power), plotted against the signal strength parameter $c$. The plots show the average rejection probability with standard error bars computed over 1000 trials for the CPT and CRT. The tests are run at level $\alpha = 0.05$.

### 6.1.2 Simulations under the alternative

Our final simulation concerns the power of the tests. Here we generate $Z$ as before, and generate $X|Z \sim \mathcal{N}(b^TZ, 1)$, exactly according the assumed distribution $Q(\cdot|Z)$, so that both tests have the nominal Type I error level $\alpha = 0.05$. Unlike the null setting, we now generate $Y|X, Z \sim \mathcal{N}(a^TZ + cX, 1)$. The strength of the signal is controlled by the parameter $c \geq 0$, where $c = 0$ corresponds to the null hypothesis being true while larger values of $c$ move farther away from the null. The results, shown in Figure 3, reveal that the CPT is slightly less powerful than the CRT across a range of values of $c$, but overall shows fairly similar performance. Thus there is only a small price to pay for the additional robustness of the CPT.

### 6.2 Capital bikeshare data set

We next implement the CPT and CRT on the Capital Bikeshare data set.\footnote{Data obtained from \url{https://www.capitalbikeshare.com/system-data}.} Capital Bikeshare is a bike sharing program in Washington, D.C., where users may check out a bike from one of their locations and return at any other location. The data set contains each ride ever taken, recording the start time and location, end time and location, bike ID number, and a user type which can be “Member” (i.e. purchasing a long-term membership in the system) or “Casual” (i.e. paying for one-time rental or a short-term pass). We use the following data:
• Test data set: all rides taken on weekdays (Monday through Friday) in October 2011. Sample size \( n = 7,346 \) rides, after an initial screening step (details below).

• Training data set (for fitting the conditional distribution \( Q(\cdot|Z) \) ): all rides taken on weekdays in September 2011 and November 2011. Sample size \( n_{\text{train}} = 149,912 \) rides.

In our experiments, we are interested in determining whether the duration, \( X \), of the ride is dependent on various factors \( Y \), such as user type (“Member” or “Casual”). Of course, the duration of the ride will be heavily dependent on the length of the route, in addition to other factors, and so to control for this we let \( Z \) encode both the route, i.e. the start and end locations, as well as the time of day at the start of the ride, since varying traffic might also affect the speed of the ride.

In order to implement the CPT and CRT, we will use a conditional normal distribution, i.e. \( (X \mid Z = z) \sim \mathcal{N}(\mu(z), \sigma^2(z)) \) as an estimate \( Q(\cdot|z) \) of \( Q_\star(\cdot|z) \). Before running the CPT or CRT, as an initial screening step we discard any test points for which we do not have a good estimate of the conditional distribution of \( X \), keeping only those test data points where we have ample training data for rides taken along the same route and at similar times of day. The details for fitting \( Q(\cdot|Z) \), and for this initial screening step, are given in Appendix B. For both the CPT and CRT, we sample \( M = 1000 \) copies of \( X \) to produce the p-value. For the CPT, the Monte Carlo sampler given in Algorithm 2 is run with \( S = 50 \) as the number of steps for producing each copy.

**Results** We test the null hypothesis \( H_0 : X \perp \!\!\!\perp Y \mid Z \) for several different choices of the response \( Y \):

• User type (“Member” or “Casual”). We might expect that “Casual” users, who are likely to be tourists or infrequent bike riders, may ride at a slower speed.

• Date, treated as continuous. Since the test data set is taken from the single month October 2011, the date of this month is a continuous variable that acts as a proxy for factors such as weather and the time of sunrise and sunset.

• Day of the week (Monday through Friday), treated as categorical. Bike riders’ behavior may differ on different days of the week, for instance, if rides on Friday are more likely to be leisure rides than the other days of the week.

For user type and date, the statistic \( T(X, Y, Z) \) that we use is the correlation between the vector \( Y \), and the vector of ride duration residuals after controlling for the effects of \( Z \)—in other words, the vector with entries \( R_i = X_i - \mathbb{E}_{X \sim Q(\cdot|Z_i)} [X] \). For day of the week, our statistic \( T(X, Y, Z) \) is given by

\[
\max_{y \in \{\text{Monday}, \ldots, \text{Friday}\}} \left| \text{Correlation between } (R_1, \ldots, R_n) \text{ and } (\mathbb{1} \{Y_1 = y\}, \ldots, \mathbb{1} \{Y_n = y\}) \right|.
\]
Table 1: p-values obtained from the CPT and CRT for the Capital Bikeshare data. The mean p-value and standard error are calculated from 10 trials of each experiment (the randomness comes from the construction of the copies $X^{(m)}$ for each test).

| Variable   | CPT p-value (std. err.) | CRT p-value (std. err.) |
|------------|-------------------------|-------------------------|
| User type  | 0.0010 (0.0000)         | 0.0010 (0.0000)         |
| Date       | 0.1146 (0.0032)         | 0.1293 (0.0032)         |
| Day of week| 0.1980 (0.0037)         | 0.2063 (0.0032)         |

Table 1 shows the resulting p-values for each choice of the variable $Y$. We can see that the CPT and CRT produce nearly identical p-values in all three cases. We conclude that the user type and duration of ride are dependent, even after controlling for our various confounding variables; on the other hand there is insufficient evidence to reach the same conclusion for the corresponding tests for the date and the day of the week.

7 Discussion

In this work, we have developed a conditional permutation test that modifies the standard permutation test of independence between $X$ and $Y$ in order to account for a known dependence of $X$ on potentially relevant confounding variables $Z$. Our theoretical results prove finite-sample Type I error control, even when the distribution of $X | Z$ is not known exactly.

We have shown that, empirically, resampling from the set of observed $X$ values preserves better Type I error control under mild errors in our model, and does not lose much power, in settings where we use intuitive statistics such as correlation between $Y$ and $X$ after regressing out the effects of $Z$. In contrast, our theoretical understanding of Type I error control covers the worst-case scenario over all possible statistics, and it may be the case that the simple statistics used in practical analyses suffer much less inflation of the Type I error. We hope to bridge this gap in future work, and also to provide some theoretical insight into the power of the CPT method, as well as to study the efficiency of the Monte Carlo sampler for the CPT and examine whether proposing swaps non-uniformly may improve the speed at which we can obtain copies $X^{(m)}$ that are not too correlated with each other.

Furthermore, in many applications it might not be possible to estimate the conditional distribution of $X | Z$ independently of the data—if only a small labeled data set $(X, Y, Z)$ is available, with no additional unlabeled data $(X, Z)$ with which to estimate this distribution, we would of course have the option of splitting the data set to use one half for fitting $Q(X|Z)$ and the remaining half to run the CPT, but this would incur substantial loss of both Type I error control and power when the sample size is limited. It is therefore important to consider how the CPT (and the CRT) can retain
their validity when the data is used for estimating $Q(X|Z)$ and then reused for testing $H_0 : X \perp Y | Z$. It is possible that tools from the selective inference literature may allow us to develop theory towards addressing this question.

Finally, both the CPT and the CRT are based in a setting where it is assumed that modeling $X | Z$ is easy while modeling $Y | X, Z$ is hard—that is, our estimate $Q(\cdot | Z)$ of the conditional distribution $X | Z$ is assumed to be highly accurate, but testing $H_0 : X \perp Y | Z$ is a substantial challenge. In contrast, many of the asymptotic tests described in Section 2.3 treat the $X$ and $Y$ variables symmetrically when testing $X \perp Y | Z$. Are there settings in which we can construct methods offering finite-sample guarantees in the style of the CPT and CRT while taking a more symmetric approach to this testing problem?

## A Proofs

### A.1 Proving validity of the CPT

**Proof of Theorem 1.** It is helpful to define some new notation to make it easier to work with permutations of our data. Let $\Pi$ be the permutation of the indices $[n] = \{1, \ldots, n\}$ given by the ranks of $X$, so that if $X_i = X_{(j)}$ then we let $\Pi(i) = j$. Under the null hypothesis that $X \perp Y | Z$, the distribution of the ranks $\Pi$ conditional on $Y, Z$ as well as on the order statistics $X()$ is given by

$$
P \{ \Pi = \pi | X(), Y, Z \} = \frac{\prod_{i=1}^{n} q(X_{(\pi(i))} | Z_i)}{\sum_{\pi'} \prod_{i=1}^{n} q(X_{(\pi'(i))} | Z_i)},$$

where the summation is taken over all $n!$ possible permutations $\pi'$.

Analogously, let $\Pi^{(m)}$ be the permutation chosen to construct the vector $X^{(m)}$, so that the sampling scheme (3) is equivalent to drawing each $\Pi^{(m)}$ according to the distribution

$$
P \{ \Pi^{(m)} = \pi \} \propto \prod_{i=1}^{n} q(X_{(\pi(i))} | Z_i).$$

Comparing this with the calculations above, we see that $\Pi^{(m)}$ is identically distributed to $\Pi$, conditional on $X(), Y, Z$. Writing $(X, Y, Z) = (X_{[\Pi]}, Y, Z)$ and $(X^{(m)}, Y, Z) = (X_{[\Pi^{(m)}]}, Y, Z)$ for each $m = 1, \ldots, M$, we therefore see that these $M + 1$ triples are i.i.d. conditional on $X(), Y, Z$. Marginalizing over $X(), Y, Z$, the triples are therefore exchangeable. $\square$

5If the unlabeled values $X_{(i)}$ are not unique, then formally, we define $\Pi$ by choosing it uniformly at random from the set of all permutations that satisfy this condition.
A.2 Proving validity of the sampling mechanisms

Proof of Theorem 2. This proof consists of simply checking the detailed balance equations for the Markov chain defined by the algorithm.

Let \( \mathcal{P} \) be the set of all partitions of \([n]\) into \([n/2]\) disjoint pairs. For any \( p \in \mathcal{P} \) and any permutations \( \pi, \pi' \), we write \( \pi \sim_p \pi' \) if \( \pi \) can be transformed to \( \pi' \) by swapping any subset of the pairs in the partition \( p \). For example, if \((i, j), (k, \ell)\) are two of the disjoint pairs in the partition \( p \), and \( \pi \) and \( \pi' \) are related via \( \pi' = \pi \circ \sigma_{ij} \circ \sigma_{k\ell} \), then \( \pi \sim_p \pi' \) (recall that \( \sigma_{ij} \) is the permutation that swaps \( i \) and \( j \)). We note that \( \sim_p \) defines an equivalence relation on the set of permutations.

We now compute the transition probability matrix of the Markov chain defined by Algorithm 1. For ease of notation, for the remainder of this proof, we will condition on \( X \). Thus, for any \( \pi, \pi' \), we have

\[
P \left\{ \Pi^t = \pi' \mid \Pi^{t-1} = \pi, \ t\text{th partition} = p \right\} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} P \left\{ \Pi^t = \pi' \mid \Pi^{t-1} = \pi, \ t\text{th partition} = p \right\},
\]

since at each time \( t \), Algorithm 1 begins by drawing a partition \( p \in \mathcal{P} \) uniformly at random. Next, given \( p \) and \( \Pi^{t-1} = \pi \), \( \Pi^t \) must satisfy \( \Pi^t \sim_p \pi \) by definition of the next step of the algorithm which can only swap pairs of indices in the partition \( p \). By examining the odds ratio defined for each \( B_{i,k} \) in (7), we see that for any \( \pi', \pi'' \sim_p \pi \),

\[
P \left\{ \Pi^t = \pi' \mid \Pi^{t-1} = \pi, \ t\text{th partition} = p \right\} = \frac{\prod_i q(X_{(\pi'(i))}|Z_i)}{\prod_i q(X_{(\pi''(i))}|Z_i)} \frac{P \{ \Pi = \pi' \}}{P \{ \Pi = \pi'' \}},
\]

where in the last step we refer to the distribution \( P \) of the permutation \( \Pi \) conditional on \( X_0, Y, Z \). Therefore,

\[
P \left\{ \Pi^t = \pi' \mid \Pi^{t-1} = \pi \right\} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{\pi''} \mathbb{1} \{ \pi' \sim_p \pi \} \cdot P \{ \Pi = \pi' \} \cdot P \{ \Pi = \pi'' \}.
\]

Thus, for any \( \pi, \pi' \), since \( \sim_p \) forms an equivalence relation over permutations, we have

\[
P \{ \Pi = \pi \} \cdot P \left\{ \Pi^t = \pi' \mid \Pi^{t-1} = \pi \right\} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} P \{ \Pi = \pi' \} \cdot P \{ \Pi = \pi'' \} \cdot \sum_{\pi''} \mathbb{1} \{ \pi' \sim_p \pi \} \cdot \mathbb{1} \{ \pi'' \sim_p \pi \} \cdot P \{ \Pi = \pi'' \}.
\]

\[
= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} P \{ \Pi = \pi' \} \cdot \mathbb{1} \{ \pi \sim_p \pi' \} \cdot P \{ \Pi = \pi \} \cdot P \{ \Pi = \pi'' \} \cdot \sum_{\pi''} \mathbb{1} \{ \pi'' \sim_p \pi' \} \cdot P \{ \Pi = \pi'' \}.
\]

\[
= P \{ \Pi = \pi' \} \cdot P \left\{ \Pi^t = \pi \mid \Pi^{t-1} = \pi' \right\}.
\]
This verifies the detailed balance equations, and so the Markov chain is reversible and has stationary distribution given by (9). Finally, it is trivial to see that this Markov chain is aperiodic and irreducible, and so this stationary distribution is unique. \(\square\)

**Proof of Theorem 3.** This result follows directly from the fact that the Markov chain defined in Algorithm 1 is reversible. This means that, under \(H_0\), the permutations \(\Pi, \Pi_\sharp, \Pi^{(1)}, \ldots, \Pi^{(M)}\) can equivalently be drawn as follows: first draw \(\Pi_\sharp\) from the distribution (9) conditional on \(X_0, Y, Z\), then draw \(\Pi, \Pi^{(1)}, \ldots, \Pi^{(M)}\) via \(M + 1\) independent runs of Algorithm 1 for \(S\) steps initialized at \(\Pi[0] = \Pi_\sharp\). Thus \(\Pi, \Pi^{(1)}, \ldots, \Pi^{(M)}\) are i.i.d. conditional on \(\Pi_\sharp, X_0, Y, Z\), and are therefore exchangeable. \(\square\)

### A.3 Proving robust Type I error control

**Proof of Theorem 4.** First we prove the result for the CRT. Let \(\tilde{X}\) be an additional copy drawn also from \(Q_n(\cdot|Z)\), independently of \(Y\) and of \(X, X^{(1)}, \ldots, X^{(M)}\). Then, since conditional on \(Y, Z\) the copies \(X, \tilde{X}, X^{(1)}, \ldots, X^{(M)}\) are independent, we have

\[
\text{d}_{TV}\left(\left((X, X^{(1)}, \ldots, X^{(M)}) \mid Y, Z\right), \left((\tilde{X}, X^{(1)}, \ldots, X^{(M)}) \mid Y, Z\right)\right) = \text{d}_{TV}\left((X \mid Y, Z), (\tilde{X} \mid Y, Z)\right) = \text{d}_{TV}(Q_n^\ast(\cdot|Z), Q_n^\ast(\cdot|Z)).
\]

Now let \(A_\alpha \subseteq (\mathcal{X})^{M+1}\) be defined as

\[
A_\alpha := \left\{(x, x^{(1)}, \ldots, x^{(M)}) \in (\mathcal{X})^{M+1} \colon \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(x^{(m)}, Y, Z) \geq T(x, Y, Z)\}}{1 + M} \leq \alpha\right\},
\]

i.e. the set where we would obtain a p-value \(p \leq \alpha\). Then

\[
P\{p \leq \alpha \mid Y, Z\} = P\{(X, X^{(1)}, \ldots, X^{(M)}) \in A_\alpha \mid Y, Z\} \\
\leq P\{(\tilde{X}, X^{(1)}, \ldots, X^{(M)}) \in A_\alpha \mid Y, Z\} \\
+ \text{d}_{TV}\left(\left((X, X^{(1)}, \ldots, X^{(M)}) \mid Y, Z\right), \left((\tilde{X}, X^{(1)}, \ldots, X^{(M)}) \mid Y, Z\right)\right) \\
= P\{(\tilde{X}, X^{(1)}, \ldots, X^{(M)}) \in A_\alpha \mid Y, Z\} + \text{d}_{TV}(Q_n^\ast(\cdot|Z), Q_n^\ast(\cdot|Z)).
\]

Finally, since \(\tilde{X}, X^{(1)}, \ldots, X^{(M)}\) are clearly i.i.d. after conditioning on \(Y, Z\), and are therefore exchangeable, by definition of \(A_\alpha\) we must have

\[
P\{(\tilde{X}, X^{(1)}, \ldots, X^{(M)}) \in A_\alpha \mid Y, Z\} \leq \alpha,
\]

proving the desired bound for the CRT.
Next we turn to the CPT, for which the analysis is more complicated since the \(X^{(m)}\)'s depend on the observed values in the vector \(X\). We will use the fact that,

For any \((U, V)\) and \((U', V')\), if \((V \mid U = u) \overset{d}{=} (V' \mid U' = u)\) for any \(u\),
then \(d_{TV}((U, V), (U', V')) = d_{TV}(U, U')\).

Let \(\tilde{X}\) be drawn from \(Q(\cdot \mid Z)\), independently of \(Y\), and let \(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\) be draws from the CPT when we sample from the values of \(\tilde{X}\) instead of \(X\). That is, independently for each \(m = 1, \ldots, M\), we draw

\[
\mathbb{P}\{\tilde{X}^{(m)} = \tilde{X}_{(\pi)} \mid \tilde{X}_{()} , Y, Z\} \propto q^n(\tilde{X}_{(\pi)} \mid Z),
\]
where \(\tilde{X}_{()}\) and \(\tilde{X}_{(\pi)}\) are defined analogously to \(X_{()}\) and \(X_{(\pi)}\) from Section 3.

Next, we observe that the \(\tilde{X}^{(m)}\)'s, conditional on \(\tilde{X}\), are generated with the same mechanism as the \(X^{(m)}\)'s conditional on \(X\). In other words, for any \(x \in \mathcal{X}^n\), we have

\[
\left(\left(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\right) \mid \tilde{X} = x, Y, Z\right) \overset{d}{=} \left(\left(X^{(1)}, \ldots, X^{(M)}\right) \mid X = x, Y, Z\right).
\]

Applying (10), then, we have

\[
d_{TV}\left(\left(\left(X, X^{(1)}, \ldots, X^{(M)}\right) \mid Y, Z\right), \left(\left(\tilde{X}, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\right) \mid Y, Z\right)\right)
\]

\[
= d_{TV}\left((X \mid Y, Z), (\tilde{X} \mid Y, Z)\right) = d_{TV}(Q^*_n(\cdot \mid Z), Q^n(\cdot \mid Z)).
\]

From this point on, we proceed as for the CRT—we have

\[
\mathbb{P}\{p \leq \alpha \mid Y, Z\} \leq \mathbb{P}\{\left(\tilde{X}, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\right) \in A_{\alpha} \mid Y, Z\} + d_{TV}(Q^*_n(\cdot \mid Z), Q^n(\cdot \mid Z)),
\]

and since \(\tilde{X}, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\) are exchangeable after conditioning on \(Y, Z\), we see that \(\mathbb{P}\{\left(\tilde{X}, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}\right) \in A_{\alpha} \mid Y, Z\} \leq \alpha\), proving the desired bound for the CPT. \(\square\)

**Proof of Theorem 5** For convenience we will write

\[
d_{TV} = d_{TV}(Q^*_n(\cdot \mid Z), Q^n(\cdot \mid Z))
\]

throughout this proof. First, by a standard property of the total variation distance, there exists a subset \(A(Z) \subseteq \mathcal{X}^n\) such that

\[
\mathbb{P}_{Q^*_n(\cdot \mid Z)}\{X \in A(Z) \mid Z\} = \mathbb{P}_{Q^n(\cdot \mid Z)}\{X \in A(Z) \mid Z\} + d_{TV}.
\]

Fix any \(M \geq 2\), and define

\[
\alpha_0(Z) := \mathbb{P}_{Q^n(\cdot \mid Z)}\{X \in A(Z) \mid Z\}, \quad \alpha(Z) := \alpha_0(Z) + 0.5 \sqrt{\log(M) \over M}.
\]
Now, by definition of the setting and the CRT, we know that conditional on \( Z \), we have \( X \sim Q^\ast_n(\cdot | Z) \) and independently, \( X^{(1)}, \ldots, X^{(M)} \sim Q^n(\cdot | Z) \). Therefore,
\[
\left( \mathbb{1} \{ X \in A(Z) \} \mid Y, Z \right) \sim \text{Bernoulli}(\alpha_0(Z) + d_{\text{TV}}),
\]
and independently,
\[
\left( \sum_{m=1}^M \mathbb{1} \{ X^{(m)} \in A(Z) \} \mid Y, Z \right) \sim \text{Binomial}(M, \alpha_0(Z)).
\]

We will work with the statistic \( T(X, Y, Z) = \mathbb{1} \{ X \in A(Z) \} \). We have
\[
P \left\{ p \leq \alpha(Z) \mid Y, Z \right\}
= P \left\{ \frac{1 + \sum_{m=1}^M \mathbb{1} \{ T(X^{(m)}, Y, Z) \geq T(X, Y, Z) \}}{1 + M} \leq \alpha(Z) \mid Y, Z \right\}
\geq P \left\{ X \in A(Z) \text{ and } \sum_{m=1}^M \mathbb{1} \{ X^{(m)} \in A(Z) \} \leq \alpha(Z) \cdot (M + 1) - 1 \mid Y, Z \right\}
= \left( \alpha_0(Z) + d_{\text{TV}} \right) \cdot P \left\{ \text{Binomial}(M, \alpha_0(Z)) \leq \alpha(Z) \cdot (M + 1) - 1 \mid Z \right\}
\geq \alpha(Z) + d_{\text{TV}} - 0.5 \sqrt{\frac{\log(M)}{M}} - P \left\{ \text{Binomial}(M, \alpha_0(Z)) > \alpha(Z) \cdot (M + 1) - 1 \mid Z \right\},
\tag{11}
\]
where the last step holds by definition of \( \alpha(Z), \alpha_0(Z) \), and the fact that \( \alpha_0(Z) + d_{\text{TV}} \leq 1 \).

Finally, it suffices to bound this binomial probability. By Bennett’s inequality, writing \( h(u) = (1 + u) \log(1 + u) - u \), for any \( t \in [0, 1] \) we have
\[
P \left\{ \text{Binomial}(M, t) > \left( t + 0.5 \sqrt{\frac{\log(M)}{M}} \right) \cdot (M + 1) - 1 \right\}
= P \left\{ \text{Binomial}(M, t) - Mt > t + 0.5 \sqrt{\frac{\log(M)}{M}} \cdot (M + 1) - 1 \right\}
\leq \exp \left\{ -Mt(1 - t) \cdot h \left( \frac{t + 0.5 \sqrt{\frac{\log(M)}{M}} \cdot (M + 1) - 1}{Mt(1 - t)} \right) \right\}
\leq \exp \left\{ -\frac{M}{4} h \left( \frac{0.5 \sqrt{\frac{\log(M)}{M}} \cdot (M + 1) - 1}{M/4} \right) \right\},
\tag{12}
\]
where the last step holds since \( h \) is an increasing function, while \( c \mapsto c \cdot h(a/c) \) is decreasing in \( c > 0 \), for any \( a > 0 \), and \( t(1 - t) \leq 1/4 \).
Finally, as $\epsilon \to 0$, we have $h(\epsilon) = \epsilon^2/2 + O(\epsilon^3)$, so as $M \to \infty$ we have

$$
\exp\left\{-\frac{M}{4} h\left(\frac{0.5 \log(M)}{M/4} \cdot (M + 1) - 1\right)\right\} = \exp\left\{-\frac{1}{2} \log(M) + o(1)\right\} = \frac{1}{\sqrt{M}} = o(1) \cdot 0.5 \sqrt{\frac{\log(M)}{M}}.
$$

Returning to (11), we see that

$$
P\{p \leq \alpha(Z) \mid Y, Z\} \geq \alpha(Z) + d_{TV} - \sqrt{\frac{\log(M)}{M}} \cdot 0.5(1 + o(1)).
$$

More concretely, for any $M \geq 2$ we can verify numerically that the quantity in (12) is bounded by $2\sqrt{\frac{\log(M)}{M}}$, which shows that the term $0.5(1 + o(1))$ above can be replaced with 2.5 for any $M \geq 2$.

## B Details for bikeshare data experiment

We will write $Z = (Z_{\text{route}}, Z_{\text{time}})$, where the route encodes both the start and end locations and is treated as categorical.

To estimate a conditional distribution $Q(\cdot \mid Z)$, we assume that $X \mid Z$ is normally distributed, and we fit the conditional mean and variance on the training data by grouping rides according to their route and taking a Gaussian kernel over their start time: for any $z = (z_{\text{route}}, z_{\text{time}})$,

$$
\hat{\mu}(z) = \sum_i \frac{w(z, Z_i^{\text{train}})}{\sum_{i'} w(z, Z_{i'}^{\text{train}})} \cdot X_i^{\text{train}}, \quad \hat{\sigma}^2(z) = \sum_i \frac{w(z, Z_i^{\text{train}})}{\sum_{i'} w(z, Z_{i'}^{\text{train}})} \cdot (X_i^{\text{train}})^2 - (\hat{\mu}(z))^2,
$$

where the weights are given by grouping observations by route and applying a Gaussian kernel to the time, i.e.

$$
w(z, Z_i^{\text{train}}) = \mathbb{1}\{(Z_i^{\text{train}})_{\text{route}} = z_{\text{route}}\} \cdot \exp\left\{-((Z_i^{\text{train}})_{\text{time}} - z_{\text{time}})^2/(2h^2)\right\}
$$

for a bandwidth $h$ of 20 minutes. Time of day is on a continuous 24 hour clock, that is, if $z_{\text{time}} = 11:00\text{pm}$ and $(Z_i^{\text{train}})_{\text{time}} = 1:00\text{am}$ then the difference between them is two hours, not 22 hours.

Our conditional distribution estimate $Q(\cdot \mid Z)$ is then given by

$$
(X \mid Z = z) \sim \mathcal{N}\left(\hat{\mu}(z), \hat{\sigma}^2(z)\right).
$$
However, since the popularity of various routes and different times of day varies widely, there are some values $z$ where our estimate of the conditional mean and variance of $X$ is unreliable due to scarce data. To check this, for any $z$ we define

$$N(z) = \sum_i w(z, Z_{i\text{train}}),$$

where a larger $N(z)$ means that there are a larger number of rides in the training data that were taken along the same route $z_{\text{route}}$, and at a time of day similar to $z_{\text{time}}$. For the test data, we then keep only those data points $(X_i, Y_i, Z_i)$ for which $N(Z_i) \geq 20$. Since this screening step uses the value of $Z_i$ but not the value of $X_i$, the $X_i$’s are still unobserved even after screening, and their distribution conditional on $Z_i$ is unchanged; therefore the CPT and CRT tests are valid even on this screened data.

**Acknowledgements**

R.F.B. was partially supported by the NSF via grant DMS-1654076 and by an Alfred P. Sloan fellowship. T.B.B. and R.J.S. were supported by an EPSRC Programme grant. R.J.S. was also supported by an EPSRC Fellowship and a grant from the Leverhulme Trust. The authors would like to thanks the Isaac Newton Institute for Mathematical Sciences for its hospitality during the programme Statistical Scalability which was supported by EPSRC Grant Number: LNAG/036, RG91310. The authors thank Samir Khan for help implementing code for our algorithms.

**References**

[1] Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.

[2] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[3] Wicher Pieter Bergsma. *Testing conditional independence for continuous random variables*. Eurandom, 2004.

[4] Thomas B Berrett and Richard J Samworth. Nonparametric independence testing via mutual information. *arXiv preprint arXiv:1711.06642*, 2017.

[5] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
[6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

[8] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. *Uncertainty In Artificial Intelligence*, 30:132–141, 2014.

[9] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 20:489–496, 2008.

[10] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[11] Julie Josse and Susan Holmes. Tests of independence and beyond. *arXiv preprint arXiv:1307.7383*, 2014.

[12] Ivan Kojadinovic and Mark Holmes. Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154, 2009.

[13] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

[14] Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

[15] Jakob Runge. Conditional independence testing based on a nearest–neighbor estimator of conditional mutual information. *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 84:938–947, 2018.

[16] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2955–2965, 2017.

[17] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

[18] Kyungchul Song. Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics*, 37(6B):4011–4045, 2009.
[19] Stephen M Stigler. Francis Galton’s account of the invention of correlation. *Statistical Science*, pages 73–79, 1989.

[20] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *arXiv preprint arXiv:1702.03877*, 2017.

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

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

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

[24] Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

[25] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, pages 2769–2794, 2007.

[26] Luca Weihs, Mathias Drton, and Nicolai Meinshausen. Symmetric rank covariances: a generalised framework for nonparametric measures of dependence. *Biometrika, to appear*, 2018.

[27] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.