ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 1, *pp.* 175–197

The conditional permutation test for independence while controlling for confounders

Thomas B. Berrett,

University of Cambridge, UK

Yi Wang and Rina Foygel Barber

University of Chicago, USA

and Richard J. Samworth

University of Cambridge, UK

[Received July 2018. Final revision September 2019]

Summary. 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 test permutes entries of X non-uniformly, to respect the existing dependence between X and Z and thus to account for the presence of these confounders. Like the conditional randomization test of Candès and co-workers in 2018, our test relies on the availability of an approximation to the distribution of $X|Z$—whereas their 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 $X|Z$, 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.

Keywords: Conditional independence testing; Exchangeability; Model misspecification; Model-X knockoffs; Permutation tests; Semisupervised learning

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 (Stigler, 1989), many measures of dependence have been proposed, including mutual information, the Hilbert–Schmidt independence criterion and distance covariance (Cover and Thomas, 2012; Gretton *et al.*, 2005; Székely *et al.*, 2007); see also Josse and Holmes (2013) for an overview. Simultaneously, much research effort has gone into developing several different tests of independence, e.g. based on ranks, kernel methods, copulas and nearest neighbours (Weihs *et al.*, 2018; Pfister *et al.*, 2018; Kojadinovic and Holmes, 2013; Berrett and Samworth, 2019). Permutation tests are particularly attractive because of their simplicity and

Address for correspondence: Thomas B. Berrett, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK.
E-mail: t.berrett@statslab.cam.ac.uk

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 of primary interest (Dawid, 1979). 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 0 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 shall study the general problem of testing $X \perp\!\!\!\perp Y | Z$. (In 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 0 after controlling for the remaining features $X^{-j} = (X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p)$; thus X^j and X^{-j} correspond to our X and Z respectively.) We are typically interested in the setting where X and Y are one dimensional whereas 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 recent work of Shah and Peters (2019) 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 proved that there is no conditional independence test that both

(a) controls type I error over any null distribution (i.e. any distribution of (X, Y, Z) with $X \perp\!\!\!\perp Y | Z$) and
(b) has better than random power against even one alternative hypothesis.

Our work complements this fundamental result of Shah and Peters (2019) by demonstrating that, given some additional knowledge, namely an approximation to the conditional distribution of X given Z, we can 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.* (2018). 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 π 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 modelling assumptions, proving that the type I error is only slightly inflated in both tests when our available distributional information is only approximately correct. In fact, in the worst case, the type I error is always *less* inflated for the new CPT method compared with 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 shall assume that the data consist of independent and identically distributed (IID) data points $(X_i, Y_i, Z_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ for $i = 1, \ldots, n$ and write $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

2.1. Permutation tests

One key reason why handling conditional independence in non-parametric contexts is so challenging is that the permutation approaches that are so effective for testing unconditional independence, $X \perp\!\!\!\perp Y$, cannot be directly applied when we seek to test conditional independence, $X \perp\!\!\!\perp Y | Z$. This is because it may be that the null hypothesis $H_0 : X \perp\!\!\!\perp Y | Z$ is true, but X and Y are highly marginally dependent because of correlation induced via each variable's dependence on Z. Under this null, if we sample a permutation π 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$, because of 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 π that permutes the X_is within the set of indices $\{i : Z_i = 0\}$ and separately permutes the X_is 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 $H_0 : X \perp\!\!\!\perp Y | Z$ does not imply that $X \perp\!\!\!\perp Y | (Z \in \text{bin } b)$; the best that 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.

Apart from independence testing, permutation tests are also popular in other settings in which the null hypothesis is exchangeable (Ernst, 2004). Moreover, Roach and Valdar (2018) developed a theory of generalized permutation tests, primarily in the context of testing simple hypotheses, for non-exchangeable null models where the weights that are assigned to permutations are non-uniform.

2.2. The conditional randomization test

The CRT, which was proposed by Candès *et al.* (2018), 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 semisupervised learning settings where unlabelled data (X, Z) are easier to obtain than labelled data (X, Y, Z), it may be possible to obtain a very accurate estimate of the conditional distribution of $X|Z$, but testing for independence with Y remains challenging because of limited sample size of the labelled data. Candès *et al.* (2018), section 1.3, gave examples of applications where unlabelled (X, Z) data are amply available whereas labelled data (X, Y, Z) are scarce—e.g. genomewide association studies, where it is important to determine whether a particular genetic variant X affects a response Y such as disease status or some other phenotype, even after controlling for the rest of the genome, encoded in Z. Human genome data, i.e. (X, Z) data, are now plentiful, but labelled data (X, Y, Z) are expensive; if we do not know the disease status Y of the individuals in previously collected samples, we need to obtain the (X, Y, Z) samples ourselves.

Assuming then that the distribution of $X|Z$ is known (or is estimated accurately from a large sample of unlabelled data), the CRT operates by sampling a new copy of the X-values in the data set. Letting $Q(\cdot|z)$ denote the distribution of X given $Z=z$, conditionally on $Z_1,\ldots,Z_n$, the CRT draws

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

independently for each $i=1,\ldots,n$, and independently of the observed X_is and Y_is. (In the special case where X is binary, earlier work by Rosenbaum (1984) proposed a related test, which is referred to as a 'CPT' but which in fact resamples X by estimating $\mathbb{P}(X=1|Z)$ with a logistic model.)

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

$$(X|Y=y, Z=z) \stackrel{\mathrm{d}}{=} (X|Z=z) \sim Q(\cdot|z),$$

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

$$(\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{Z}) \stackrel{\mathrm{d}}{=} (\mathbf{X}, \mathbf{Y}, \mathbf{Z})$$

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

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

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

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

are therefore exchangeable as well. We can compute a p-value by ranking the value that is obtained from the true $\mathbf{X}$-vector against the values that are obtained from the CRT's copies:

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

The exchangeability of the random variables in expression (1) ensures that this is a valid p-value under the null, i.e. it satisfies $\mathbb{P}(p \leqslant \alpha) \leqslant \alpha$ for all $\alpha \in [0,1]$ if the null hypothesis H_0 is true.

The 'model-X knockoffs' framework of Candès *et al.* (2018) 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 that are used to construct the knockoff copies of each feature (analogous to the $\mathbf{X}^{(m)}$s above), was studied by Barber *et al.* (2019).

2.3. Other tests of conditional independence

Before introducing our new work, we give a brief overview of some additional conditional independence testing methods that have been proposed in the literature. Many methods assume some parametric model for the response Y, such as a linear model, $Y=\alpha X+\beta^{\mathrm{T}}Z+(\text{noise})$, 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 $Y - \hat{\beta}^{\mathrm{T}} Z$ is correlated with X. Belloni *et al.* (2014) proposed a variant on this approach, which assumes approximate linear models for both Y and X. Their method regresses both X and Y on Z and 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 (2019) 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 in reproducing kernel Hilbert spaces have been studied by, for example, Fukumizu *et al.* (2008), Zhang *et al.* (2011) and Strobl *et al.* (2019). Other works use permutations of the data, including Doran *et al.* (2014) and Sen *et al.* (2017), where the methods have the flavour of binning Z and then permuting within groups. Bergsma (2004), Song (2009) and Veraverbeke *et al.* (2011) studied copula methods for testing conditional independence. There is also a large literature on extending measures of marginal independence to the conditional setting, including partial distance covariance (Székely and Rizzo, 2014), conditional mutual information (Runge, 2018), characteristic functions (Su and White, 2007), Hellinger distances (Su and White, 2008) and smoothed empirical likelihoods (Su and White, 2014).

A related problem is that of testing the null hypothesis that a certain treatment has no effect in a randomized experiment. In the treatment effects literature it is common to calculate p-values by comparing a test statistic with null statistics that are based on randomly reassigning treatments in the data. However, in some situations, uniformly random reassignment is inappropriate and does not result in valid p-values, because of some underlying structure; see Athey *et al.* (2018) for network dependence and Hennessy *et al.* (2016) for covariate imbalance. In such cases it is sometimes possible to develop non-uniform randomization schemes that result in valid p-values, as with the CPT and the CRT.

3. The conditional permutation test

Recall that the CRT (Candès *et al.*, 2018) creates copies $\mathbf{X}^{(m)}$ of the vector $\mathbf{X}$ sampled under the null hypothesis that $X \perp\!\!\!\perp Y|Z$, by drawing

$$\mathbf{X}^{(m)}|\mathbf{X}, \mathbf{Y}, \mathbf{Z} \sim Q^n(\cdot|\mathbf{Z}), \tag{2}$$

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

Our proposed method, the CPT, is a variant on the CRT, with $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ drawn as in distribution (2) but under the constraint that each $\mathbf{X}^{(m)}$ must be a permutation of the original vector $\mathbf{X}$. Once we have drawn $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$, they will then be used exactly as for the CRT—given some predefined statistic $T = T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, our p-value is given by

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{X}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geqslant T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\}}{1 + M}. \tag{3}$$

All that remains, then, is to specify how these permuted copies $\mathbf{X}^{(m)}$ will be drawn.

To draw the $\mathbf{X}^{(m)}$s, we first need to define some notation. Let $\mathcal{S}_n$ denote the set of permutations on the indices $\{1, \ldots, n\}$. Given any vector $\mathbf{x} = (x_1, \ldots, x_n)$ and any permutation $\pi \in \mathcal{S}_n$, define $\mathbf{x}_\pi = (x_{\pi(1)}, \ldots, x_{\pi(n)})$, i.e. the vector $\mathbf{x}$ with its entries reordered according to the permutation π.

The CPT copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are then drawn as follows: after observing $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$, we draw M permutations $\pi^{(1)}, \ldots, \pi^{(M)}$ according to the conditional distribution of $X|Z$, and then apply these permutations to $\mathbf{X}$. Specifically, let

$$\mathbf{X}^{(m)} = \mathbf{X}_{\pi^{(m)}} \qquad \mathbb{P}(\pi^{(m)} = \pi | \mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \frac{q^n(\mathbf{X}_\pi | \mathbf{Z})}{\sum_{\pi' \in \mathcal{S}_n} q^n(\mathbf{X}_{\pi'} | \mathbf{Z})}. \tag{4}$$

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 ν on $\mathcal{X}$ that does not depend on z. We write $q^n(\cdot|\mathbf{Z}) := q(\cdot|Z_1) \ldots q(\cdot|Z_n)$ 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.

Why is this the right distribution for drawing the permuted copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$? To understand this, it is helpful to consider a different formulation of the permutation scheme. Let $\mathbf{X}_{()} = (X_{(1)}, \ldots, X_{(n)})$ be the order statistics of the list of values $\mathbf{X} = (X_1, \ldots, X_n)$. (In the setting where $\mathcal{X} = \mathbb{R}$, we can of course use the usual ordering on $\mathbb{R}$. In the general case we can simply take an arbitrary total ordering on $\mathcal{X}$; the choice of ordering is irrelevant as its only role is to allow us to observe the set of values of $\mathbf{X}$ without knowing which one corresponds to which data point.) Define also $\mathbf{X}_{(\pi)} = (X_{(\pi(1))}, \ldots, X_{(\pi(n))})$ for each $\pi \in \mathcal{S}_n$, and let $\Pi \in \mathcal{S}_n$ be the permutation given by the ranks of the true observed vector $\mathbf{X}$, so that $\mathbf{X} = \mathbf{X}_{(\Pi)}$. In other words, $\mathbf{X}_{()}$ gives the order statistics of $\mathbf{X}$, and Π reveals the ranks; together these two pieces of information are sufficient to reconstruct $\mathbf{X}$. (If the unlabelled values $X_{(i)}$ are not unique then, formally, we define Π by choosing it uniformly at random from the set of all permutations that satisfy this condition.)

Under the null hypothesis that $X \perp\!\!\!\perp Y | Z$, we can verify that the distribution of the true ranks Π, conditionally on $\mathbf{Y}$ and $\mathbf{Z}$ as well as on the order statistics $\mathbf{X}_{()}$, is given by

$$\mathbb{P}(\Pi = \pi | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}) = \frac{q^n(\mathbf{X}_{(\pi)} | \mathbf{Z})}{\sum_{\pi' \in \mathcal{S}_n} q^n(\mathbf{X}_{(\pi')} | \mathbf{Z})}. \tag{5}$$

Furthermore, examining the definition (4) of the CPT copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$, we can see that the CPT can equivalently be defined by

$$\mathbf{X}^{(m)} = \mathbf{X}_{(\Pi^{(m)})} \tag{6}$$

where $\Pi^{(m)} | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}$ is drawn from equation (5). In fact, comparing with expression (4), we see that $\Pi^{(m)} = \Pi \circ \pi^{(m)}$.

The following theorem formalizes the above intuition and verifies that this procedure yields a valid test of H_0.

Theorem 1. Assume that $H_0 : X \perp\!\!\!\perp Y | Z$ is true, and that the conditional distribution of $X|Z$ is given by $Q(\cdot|Z)$. Suppose that $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are drawn IID from the CPT sampling scheme (4). Then the $M + 1$ triples

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

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 expression (3) is valid, satisfying $\mathbb{P}(p \leqslant \alpha) \leqslant \alpha$ for any desired type I error rate $\alpha \in [0, 1]$ when H_0 is true.

Proof. Our work above verified that, under H_0, the true data vector $\mathbf{X}$ and the CPT copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are permutations of $\mathbf{X}_{()}$ obtained via IID draws from expression (5), conditionally

on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$. Therefore, after marginalizing over $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$, the $M+1$ triples $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, $(\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{Z}), \ldots, (\mathbf{X}^{(M)}, \mathbf{Y}, \mathbf{Z})$ are exchangeable.

3.1. Comparing the conditional permutation test and conditional randomization test

To compare the construction of the copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ in each of the two methods, for the CPT, the copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are IID draws from the null distribution of $\mathbf{X}$, conditionally on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$. In comparison, the CRT copies defined in expression (2) are IID draws from the null distribution of $\mathbf{X}$ conditioned on $\mathbf{Y}$ and $\mathbf{Z}$—but without conditioning on $\mathbf{X}_{()}$.

Each of these two constructions yields a valid test if the distribution $Q(\cdot|z)$, which is used to draw the (resampled or permuted) copies $\mathbf{X}^{(m)}$, is correct—i.e. if we know the true conditional distribution of $X|Z$. This result is proved in theorem 1 above for the CPT, whereas the analogous result for the CRT is proved in lemma F.1 on page 4 of the on-line supplement to Candès *et al.* (2018). However, if the null hypothesis is *not* true, which method might be more sensitive and better able to detect a non-null? Furthermore, what might occur for these two methods if $Q(\cdot|z)$ is not exactly correct? We next explore the difference between the two methods in greater detail to begin to address these questions.

3.1.1. Use of marginal distribution of X

In terms of how the tests are run, the difference between the CPT and CRT can be described as follows: whereas both tests use the (true or estimated) conditional distribution $Q(\cdot|Z)$, the CPT additionally uses the marginal distribution of the observed data vector $\mathbf{X}$, by observing its unlabelled values $\mathbf{X}_{()}$. Intuitively, using this additional information can in some cases make the copies $\mathbf{X}^{(m)}$ more similar to the original $\mathbf{X}$, 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 shall develop theory to examine the two tests' robustness to errors in estimating the conditional distribution $Q(\cdot|Z)$, and we shall compare the tests in terms of both type I error and power in experiments in Section 6.

3.1.2. 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 are functions $h(\cdot)$ and $c(\cdot)$ 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 $\mathcal{X}$, whereas $c(z)$ adjusts the normalizing constants as needed.)

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

$$q_1^n(\mathbf{X}_\pi|\mathbf{Z}) = \prod_{i=1}^{n} q(X_{\pi(i)}|Z_i)h(X_{\pi(i)})c(Z_i) = q^n(\mathbf{X}_\pi|\mathbf{Z}) \prod_{i=1}^{n} h(X_i)c(Z_i).$$

The first term, $q^n(\mathbf{X}_\pi|\mathbf{Z})$, is the same as for the CPT run with conditional density q, whereas the second term, $\Pi_{i=1}^n h(X_i)c(Z_i)$, does not depend on the permutation π and therefore does not affect the resulting distribution of the sampled permutations. In other words, the CPT sampling distribution (4) is unchanged if we replace q with q_1.

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—that is, theorem 1 holds whenever the conditional distribution $Q(\cdot|Z)$ has a density of the form $q(x|z)h(x)c(z)$, for some functions

$h(\cdot)$ and $c(\cdot)$. 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 $\exp\{xz^{\mathrm{T}}\theta - f(x) - g(z)\}$, in which case the CPT would need to estimate only the parametric component θ. 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 conditional permutation test

To run the CPT, we need to be able to sample permutations $\Pi^{(1)}, \ldots, \Pi^{(M)}$ from distribution (4). 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 π, draws its proposed permutation π' uniformly at random. For even a moderate n, however, the acceptance odds ratio

$$\frac{q^n(\mathbf{X}_{\pi'}|\mathbf{Z})}{q^n(\mathbf{X}_{\pi}|\mathbf{Z})} = \frac{\prod_{i=1}^n q(X_{\pi'(i)}|Z_i)}{\prod_{i=1}^n q(X_{\pi(i)}|Z_i)} \tag{7}$$

will be extremely low for nearly all permutations π' (unless, of course, the dependence of X on Z is very weak). In other words, a uniformly drawn permutation π' 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 π, we propose the permutation $\pi' = \pi \circ \sigma_{ij}$, where σ_{ij} is the permutation that swaps indices i and j, which are drawn at random. The acceptance odds ratio (7) now simplifies to

$$\frac{q(X_{\pi(j)}|Z_i)\, q(X_{\pi(i)}|Z_j)}{q(X_{\pi(i)}|Z_i)\, q(X_{\pi(j)}|Z_j)}. \tag{8}$$

The probability of accepting a swap will now be reasonably high; however, each step can alter only 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 $\lfloor n/2 \rfloor$ disjoint pairs of indices from $\{1, \ldots, n\}$. Next, independently and in parallel for each pair, we decide whether or not to swap this pair (i, j), according to the odds ratio (8). This sampler is defined formally in algorithm 1 in Table 1. For ease of our theoretical

Table 1. Algorithm I: parallelized pairwise sampler for the CPT

Input: initial permutation $\Pi^{[0]}$, integer $S \geqslant 1$
for $s = 1, 2, \ldots, S$ *do*
sample uniformly without replacement from $\{1, \ldots, n\}$ to obtain disjoint pairs

$$(i_{s,1}, j_{s,1}), \ldots, (i_{s,\lfloor n/2 \rfloor}, j_{s,\lfloor n/2 \rfloor})$$

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

$$\frac{\mathbb{P}(B_{s,k}=1)}{\mathbb{P}(B_{s,k}=0)} = \frac{q(X_{(\Pi^{[s-1]}(j_{s,k}))}|Z_{i_{s,k}})\, q(X_{(\Pi^{[s-1]}(i_{s,k}))}|Z_{j_{s,k}})}{q(X_{(\Pi^{[s-1]}(i_{s,k}))}|Z_{i_{s,k}})\, q(X_{(\Pi^{[s-1]}(j_{s,k}))}|Z_{j_{s,k}})}$$

define $\Pi^{[s]}$ by swapping $\Pi^{[s-1]}(i_{sk})$ and $\Pi^{[s-1]}(j_{sk})$ for each k with $B_{sk} = 1$
end for

analysis, we shall work with the order statistics $\mathbf{X}_{()}$, rather than the original ordered vector $\mathbf{X}$, in our sampler; this difference is only in the notation, i.e. the algorithm can equivalently be implemented with $\mathbf{X}$ in place of $\mathbf{X}_{()}$.

The next theorem verifies that the resulting Markov chain yields the desired stationary distribution. (The proof of this theorem and all remaining proofs are given in Appendix A.)

Theorem 2. For every initial permutation $\Pi^{[0]}$, the distribution (5) of the permutation Π conditionally on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$ is a stationary distribution of the Markov chain defined in algorithm 1. If additionally $q(x|z) > 0$ for all $x \in \mathcal{X}$ and all $z \in \mathcal{Z}$, then it is the unique stationary distribution.

This result justifies the thought that, if algorithm 1 is run for a sufficient number of steps S, then the resulting copy $\mathbf{X}_{(\Pi^{[S]})}$ acts as an appropriate control for $\mathbf{X}$ in testing conditional independence. In fact, though, we can make a much stronger statement—since the original permutation Π also follows distribution (5) conditionally on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$ under the null, this means that, by initializing algorithm 1 at $\Pi^{[0]} = \Pi$ (i.e. at the original data vector $\mathbf{X}$), we are initializing with a draw from the stationary distribution. Therefore $\mathbf{X}^{[S]} = \mathbf{X}_{(\Pi^{[S]})}$ is a draw from the target distribution at any S and is a valid control for $\mathbf{X}$ even if the number of steps S is small. Of course, if S is too small, then the control copy will be too similar to the original data vector $\mathbf{X}$, and our power to reject the null will be low; we explore this empirically in Section 6, and we shall see that the sampler mixes well at even a moderate S (for example, in our experiments, we used $S = 50$).

In practice, we want to draw M copies, $\mathbf{X}^{(m)}$ for $m = 1, \ldots, M$, and we need to ensure that the original data $\mathbf{X}$ and each of the M permutations $\mathbf{X}^{(m)}$ are all exchangeable with each other. If we sample the permuted vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ sequentially, by running algorithm 1 for SM steps and extracting one copy $\mathbf{X}^{(m)}$ 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 algorithm 2 in Table 2.

Algorithm 2 provides an exchangeable sampling mechanism, since the permutation $\Pi_\sharp$ is at the 'centre', lying S steps away from each of the permutations $\Pi, \Pi^{(1)}, \ldots, \Pi^{(M)}$. The following result verifies exchangeability.

Theorem 3. Let $\mathbf{X}_{()}$ and Π be the order statistics and ranks of $\mathbf{X}$, as defined previously, so that $\mathbf{X} = \mathbf{X}_{(\Pi)}$. Let $\Pi^{(1)}, \ldots, \Pi^{(M)}$ be the output of algorithm 2, when initialized at $\Pi_{\text{init}} = \Pi$, and let $\mathbf{X}^{(m)} = \mathbf{X}_{(\Pi^{(m)})}$ for each $m = 1, \ldots, M$. Assume that the null hypothesis that $X \perp\!\!\!\perp Y|Z$ holds, and the conditional distribution of $X|Z$ is given by $Q(\cdot|Z)$, so that the distribution of Π conditionally on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$ is given by expression (5). Then the triples $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{Z}), \ldots, (\mathbf{X}^{(M)}, \mathbf{Y}, \mathbf{Z})$ are exchangeable.

Table 2. Algorithm 2: exchangeable sampler for multiple draws from the CPT

Input: initial permutation Π_{init} and integer $S \geqslant 1$
Define $\Pi_\sharp$ 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 $\Pi^{[0]} = \Pi_\sharp$ for S steps
end for

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

5. Robustness of the conditional permutation test and conditional randomization test

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 $Q(\cdot|Z)$ that we use for sampling when running the CPT or CRT is only an approximation to the true conditional, denoted by $Q_*(\cdot|Z)$. 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_* and its approximation Q. Throughout, we shall assume that the statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R}$ that is used in the test as well as the approximation Q to the conditional distribution are chosen independently of $\mathbf{X}$ and $\mathbf{Y}$. For instance, in many applications, we may have access to unlabelled data, i.e. draws of (X, Z) without Y, which we can use to construct an estimate Q.

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

Theorem 4. Assume that $H_0 : X \perp\!\!\!\perp Y|Z$ is true, and that the conditional distribution of $X|Z$ is given by $Q_*(\cdot|Z)$. For a fixed integer $M \geqslant 1$, let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ be copies of $\mathbf{X}$ generated either from the CRT (2) from the CPT (4) or from the exchangeable sampler for the CPT (algorithm 2) with any fixed parameter $S \geqslant 1$, using an estimate Q of the true conditional distribution Q_*.

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

$$\mathbb{P}(p \leqslant \alpha | \mathbf{Y}, \mathbf{Z}) \leqslant \alpha + d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\},$$

where p is the p-value that is computed in expression (3), and the probability is taken with respect to the distribution of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ conditionally on $\mathbf{Y}$ and $\mathbf{Z}$.

Of course, we can also bound the type I error rate unconditionally, with

$$\mathbb{P}(p \leqslant \alpha) \leqslant \alpha + \mathbb{E}[d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}],$$

which we obtain from the result above by marginalizing over $\mathbf{Y}$ and $\mathbf{Z}$.

This result ensures that, if Q is a good approximation to Q_*, then both the CPT and the 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 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 $\mathbf{X}$ and $\mathbf{Y}$, could be more robust to errors than theorem 4 suggests.

Although 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 tight when the number of copies $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ is large.

Theorem 5. Under the setting and assumptions of theorem 4, there is a statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R}$ such that, for the CRT,

$$\sup_{\alpha \in [0,1]} \{\mathbb{P}(p \leqslant \alpha | \mathbf{Y}, \mathbf{Z}) - \alpha\} \geqslant d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} - 0.5\{1 + o(1)\}\sqrt{\left\{\frac{\log(M)}{M}\right\}}$$

as $M \to \infty$. (To be more precise with the constant, we can replace $0.5\{1+o(1)\}$ with 2.5 for any $M \geqslant 2$.)

In other words, if we use the statistic T that is best able to detect errors in our conditional distribution, and we choose α adversarially, then the excess type I error of the CRT is exactly equal to $d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}$ (up to a vanishing factor), and therefore is at least as high as that of the CPT under *any* statistic.

Unlike for the CRT, we have found that there is no simple characterization of the worst-case scenario for the CPT. In particular, for some specially constructed distributions on (X, Y, Z), we can show that the CPT achieves the same lower bound as given in theorem 5 for the CRT (again, under a worst-case choice of the statistic T), but for other joint distributions on (X, Y, Z) we can verify that the CPT cannot achieve this error rate. In particular, since the CPT is invariant to the base measure (as discussed in Section 3.1), if $Q(\cdot|z)$ is correct up to the base measure, then the excess type I error of the CRT may be as large as $d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}$ whereas the CPT is guaranteed to control the type I error at level α.

It is important to note that the lower bound for the CRT in theorem 5 applies only to a specific worst-case statistic 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. However, in Section 6 we shall see that, empirically, the CPT often yields a far lower type I error than does the CRT in simulations. Thus, we interpret theorem 5 as giving us a partial theoretical understanding of this phenomenon, since it addresses only the worst-case statistic.

5.1. When is the total variation distance small?

For theorem 4 to have practical implications, we need to verify that there are settings where, although the true distribution Q_* of $X|Z$ is unknown, it can be estimated to high accuracy, with $d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} = o_p(1)$ (so that excess type I error is guaranteed to be small). As discussed in Section 2.2, in many applications we may have a large unlabelled data set, say $(X_i^{\mathrm{unlab}}, Z_i^{\mathrm{unlab}})$, $i = 1, \ldots, N$, with which we can compute an estimate Q of Q_*. As discussed by Barber and Candès (2019) in the setting of model-X knockoffs, the unlabelled data set does not need to have the same distribution over (X, Z) as the labelled data, as long as the conditional distribution of $X|Z$ is the same.)

In this section, we briefly sketch two settings where, given a large unlabelled sample size N, our estimate Q is likely to satisfy $d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} = o_p(1)$. Our results here are stated informally, with no technical details, since we aim only to give intuition for the settings where theorem 4 is useful.

5.1.1. Parametric setting

We shall use Pinsker's inequality relating total variation distance to the Kullback–Leibler divergence, namely

$$d_{\mathrm{TV}}^2\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} \leqslant \frac{1}{2} d_{\mathrm{KL}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} = \frac{1}{2} \sum_{i=1}^{n} d_{\mathrm{KL}}\{Q_*(\cdot|Z_i), Q(\cdot|Z_i)\}.$$

It is therefore sufficient to show that $\Sigma_{i=1}^n d_{\mathrm{KL}}\{Q_*(\cdot|Z_i), Q(\cdot|Z_i)\} = o_p(1)$.

If the true conditional distribution $Q_*(\cdot|z)$ belongs to a parametric family, then this will typically hold whenever the unlabelled sample size satisfies $N \gg nk$, where k is the number of parameters defining the models in the family. Specifically, we can think of a setting where $Q_*(\cdot|z)$ has density $f_{\theta_*}(\cdot|z)$, where $\theta_* \in \mathbb{R}^k$ is the unknown parameter vector while the family of densities

$f_\theta(\cdot|z)$ is known. For example, suppose that $\mathcal{Z} = \mathbb{R}^{k-1}$, and the conditional distribution of $X|Z$ is given by

$$X|Z = z \sim \mathcal{N}(z^{\mathrm{T}}\beta_*, \sigma_*^2).$$

Then the unknown parameters are $\theta_* = (\beta_*, \sigma_*^2)$ and standard least squares theory enables us to produce independent (maximum likelihood) estimates $\hat{\beta}$ and $\hat{\sigma}^2$ satisfying

$$\hat{\beta} \sim N_{k-1}\{\beta_*, \sigma_*^2(\mathbf{Z}_{\mathrm{unlab}}^{\mathrm{T}}\mathbf{Z}_{\mathrm{unlab}})^{-1}\},$$
$$\hat{\sigma}^2 \sim \frac{\sigma_*^2}{N}\chi^2_{N-k+1},$$

where $\mathbf{Z}_{\mathrm{unlab}}$ is the $N \times (k-1)$ matrix with ith row Z_i^{unlab}. Thus, for any $z \in \mathcal{Z}$,

$$\begin{aligned} d_{\mathrm{KL}}\{Q_*(\cdot|z), Q(\cdot|z)\} &= d_{\mathrm{KL}}\{\mathcal{N}(z^{\mathrm{T}}\beta_*, \sigma_*^2), \mathcal{N}(z^{\mathrm{T}}\hat{\beta}, \hat{\sigma}^2)\} \\ &= \log\left(\frac{\hat{\sigma}}{\sigma_*}\right) + \frac{\sigma_*^2}{2\hat{\sigma}^2} - \frac{1}{2} + \frac{(z^{\mathrm{T}}\hat{\beta} - z^{\mathrm{T}}\beta_*)^2}{2\hat{\sigma}^2} = O_p\left(\frac{1+\|z\|^2}{N}\right) \end{aligned}$$

under mild conditions on the distribution of Z. Putting everything together, if Z has a finite second moment we then have

$$d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} = O_p\left\{\sqrt{\left(n\frac{k}{N}\right)}\right\},$$

which is vanishing as long as the unlabelled sample size satisfies $N \gg nk$.

5.1.2. Non-parametric setting with binary data

As a second example, suppose that $\mathcal{X} = \{0, 1\}$, so that estimating $Q_*(\cdot|z)$ is equivalent to estimating the regression function $p_*(z) := \mathbb{P}(X = 1|Z = z)$. Assuming that this probability is bounded away from 0 and 1, and again applying Pinsker's inequality, we see that, under mild conditions,

$$d_{\mathrm{TV}}^2\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} \leqslant \frac{1}{2}\sum_{i=1}^n d_{\mathrm{KL}}\{Q_*(\cdot|Z_i), Q(\cdot|Z_i)\} \asymp \sum_{i=1}^n \{\hat{p}(Z_i) - p_*(Z_i)\}^2,$$

where $\hat{p}(z)$ is our estimate of $p_*(z) = \mathbb{P}(X = 1|Z = z)$ based on the unlabelled sample.

Since we are working in a non-parametric setting, suppose that we estimate $p_*(z) = \mathbb{P}(X = 1|Z = z)$ via a kernel method, working in a low dimensional space $\mathcal{Z} = \mathbb{R}^k$. Then standard non-parametric theory ensures that, for z in a high probability subset of $\mathcal{Z}$, we can achieve error

$$\{\hat{p}(z) - p_*(z)\}^2 \sim N^{-a_k},$$

where the exponent a_k is a small positive value, depending on both the ambient dimension k and the properties of the function $z \mapsto p_*(z)$ (e.g. smoothness or Lipschitz properties). Therefore, we can expect to have

$$d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\} \lesssim \sqrt{(nN^{-a_k})},$$

which is vanishing whenever the unlabelled sample size N is sufficiently large relative to the labelled sample size n.

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. The code that was used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets.`

6.1. Simulated data: power and error control

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 $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ 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 $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = |\mathrm{corr}(\mathbf{X}, \mathbf{Y})|$ as our test statistic. We generate $M = 500$ copies of $\mathbf{X}$ under either the 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 $H_0 : X \perp\!\!\!\perp Y|Z$ holds.

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

$$a, b \sim \mathcal{N}_p(0, \mathbf{I}_p).$$

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

$$\begin{aligned} Z &\sim \mathcal{N}_p(0, \mathbf{I}_p), \\ X|Z &\sim Q_*(\cdot|Z), \\ Y|X, Z &\sim \mathcal{N}(p^{-1} a^{\mathrm{T}} Z, 1), \end{aligned}$$

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

Throughout, the estimated conditional distribution of $X|Z$ will be given by

$$Q(\cdot|Z) = \mathcal{N}(b^{\mathrm{T}} Z, 1),$$

but this estimate might not be exactly correct. We shall consider several sources of error in this model.

(a) Non-linear mean: one source of error comes from assuming a linear relationship between variables where this is not so. We choose sample size $n = 50$ and try three simple examples, taking $Q_*(\cdot|z) = \mathcal{N}\{\mu(z), 1\}$, where $\mu(\cdot)$ is given by
   (i) quadratic, $\mu(z) = b^{\mathrm{T}} z + \theta (b^{\mathrm{T}} z)^2$,
   (ii) cubic, $\mu(z) = b^{\mathrm{T}} z - \theta (b^{\mathrm{T}} z)^3$, and
   (iii) hyperbolic tangent, $\mu(z) = \tanh(\theta b^{\mathrm{T}} z)/\theta$.

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

(b) Coefficients estimated on unlabelled data: even if the form of the model for $X|Z$ is correct, the coefficients b may not be known perfectly. As described earlier, in many practical settings we may have access to ample unlabelled data (X, Z), separate from our labelled data set of points (X, Y, Z) that is used to test the hypothesis of conditional independence. For this setting, we estimate the unknown coefficient vector b by $\hat{b}$, defined as the least squares estimate by using an unlabelled sample $(X_i^{\text{unlab}}, Z_i^{\text{unlab}})$, $i = 1, \ldots, N$, generated independently of the data points (X_i, Y_i, Z_i). This experiment is repeated for unlabelled sample sizes $N = 50, 100, \ldots, 500$. The labelled sample size is given by $n = 50$ in each case.

(c) Coefficients estimated by reusing the data: finally, in settings where unlabelled data may not be available, we may be tempted to estimate the model of $X|Z$ simply by using our data points (X_i, Y_i, Z_i), $i = 1, \ldots, n$. This approach is not covered by our theory (since the conditional distribution $Q(X|Z)$ is data dependent in this case), but it is certainly of practical interest to see how the method performs in this setting. We test sample sizes $n = 50, 100, \ldots, 500$, in each case estimating the unknown true coefficient vector b by $\hat{b}$, which in this case is now given by the least squares regression of X on Z trained on the *same* data set, $(X_1, Z_1), \ldots, (X_n, Z_n)$.

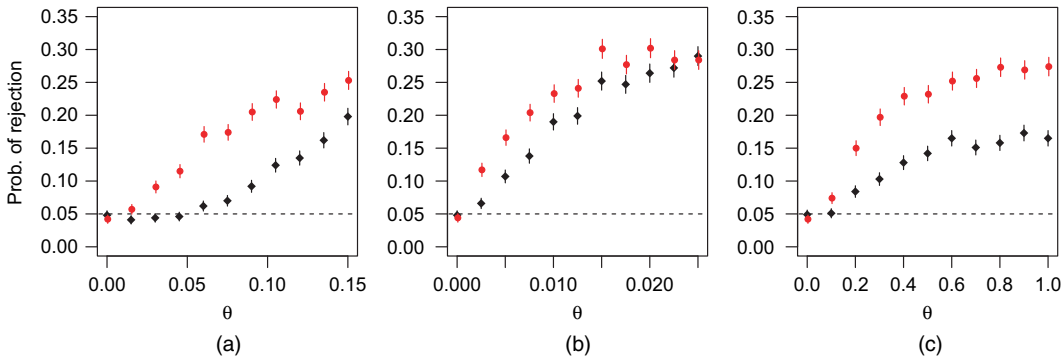

Fig. 1. Simulation results for robustness to misspecification of the mean function for (a) quadratic $\mu(z)$, (b) cubic $\mu(z)$ and (c) hyperbolic tangent $\mu(z)$: the figures show the probability of rejection (i.e. the type I error rate), plotted against the model misspecification parameter θ; the plots show the average rejection probability with standard error bars computed over 1000 trials for the CPT (◆) and CRT ● (- - - -, nominal level $\alpha = 0.05$)

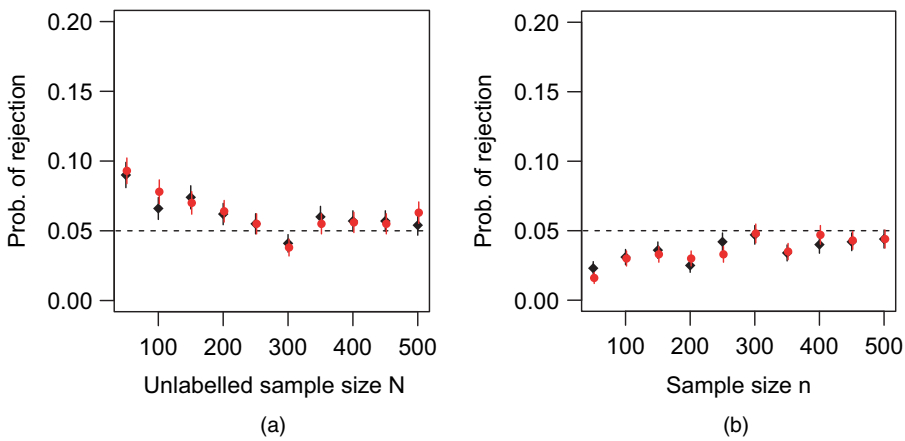

Fig. 2. Simulation results for robustness to models trained on (a) unlabelled data or (b) by reusing the data (the details are as for Fig. 1): ◆, CPT; ●, CRT

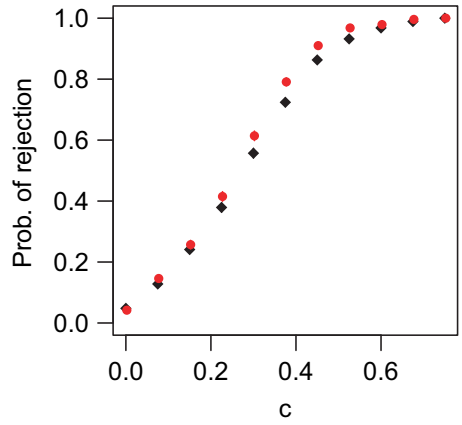

Fig. 3. Simulation results testing power under the alternative: the figure shows the probability of rejection (i.e. the power), plotted against the signal strength parameter c; the plot shows the average rejection probability with standard error bars computed over 1000 trials for the CPT (♦) and CRT (●); the tests are run at level $\alpha = 0.05$

6.1.1.1. Results. The plots in Figs 1 and 2 show the results of these experiments when we have a non-linear mean, and when we estimate the coefficients by using unlabelled data or reusing data respectively. As the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ is true in all 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 θ is not too large (for the non-linear mean setting) or when the unlabelled sample size N or labelled sample size n is not too small (when the model coefficients are trained on unlabelled data or reused data).

For the non-linear mean experiments, in Fig. 1 we see that in many cases the CPT is significantly more robust than the CRT. The cases $\theta = 0$ confirm that both tests achieve the nominal type I error level $\alpha = 0.05$ when the assumed distribution Q is correct. As the misspecification parameter θ increases (so that the model $Q(\cdot|z)$ that we use for running the CPT or CRT grows further from the true model $Q_*(\cdot|z)$), we see that both methods suffer an inflation of the type I error level, but for the CPT the excess type I error is substantially lower than that of the CRT.

Next, we turn to the setting where the estimated model $Q(\cdot|z)$ is obtained by regressing X on Z by using either a separate unlabelled data set, shown in Fig. 2(a), or by reusing the same data set, shown in Fig. 2(b). The results are encouraging, showing that, when using unlabelled data, the type I error is already very close to the nominal level as soon as the unlabelled sample size N is larger than n. When reusing the data, the method appears to be quite conservative at smaller sample sizes n—the cause of this is an interesting question that we hope to study in future work.

6.1.2. Simulations under the alternative

Our final simulation concerns the power of the tests. Here we generate Z as before, and generate $X|Z \sim \mathcal{N}(b^{\mathrm{T}}Z, 1)$, exactly according to 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 $Y|X, Z \sim \mathcal{N}(a^{\mathrm{T}}Z + cX, 1)$. The strength of the signal is controlled by the parameter $c \geqslant 0$, where $c = 0$ corresponds to the null hypothesis being true whereas larger values of c move further away from the null. The results, which are shown in Fig. 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.

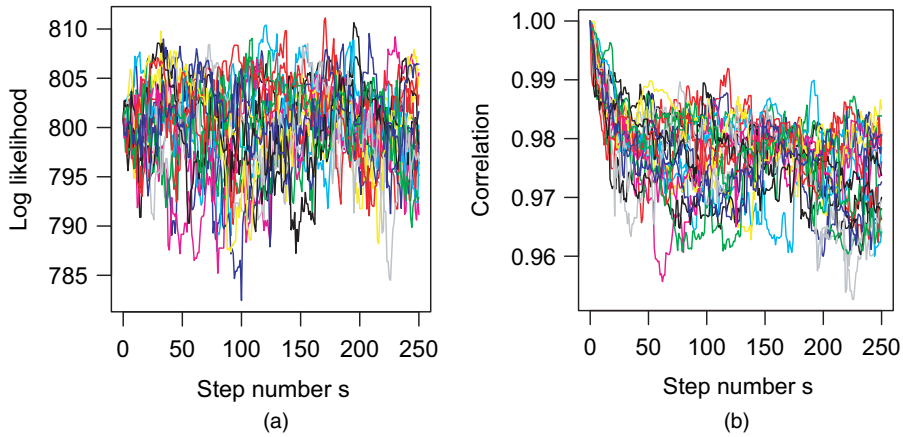

Fig. 4. Simulation results showing trace plots for the CPT sampler, examining the CPT copy $\mathbf{X}^{[s]}$ at step s of algorithm 1: (a) log-likelihood of $\mathbf{X}^{[s]}$; (b) corr$(\mathbf{X}, \mathbf{X}^{[s]})$

6.2. *Simulated data: mixing of the conditional permutation test sampler*

In practice, we cannot implement the CPT method as defined in expression (4) (unless, of course, the sample size n is so small that we can simply enumerate all $n!$ possible permutations). Instead, in our experiments, we use the exchangeable Markov chain Monte Carlo sampler, which is defined in algorithm 2. All our simulations and real data experiments implement this sampler with $S = 50$, meaning that the Markov chain is run for 50 steps for each new permuted copy $\mathbf{X}^{(m)}$ of the data. Is this moderate number of steps sufficient to ensure that the chain has mixed well, or are we producing highly correlated data that will lead to reduced power? To examine this question, we generate one data set, consisting of confounders Z and feature X generated exactly as in Section 6.1.2, and then run the parallel pairwise sampler (algorithm 1) independently for 20 trials (i.e. each time initializing at the same original data). At each iteration, setting $\mathbf{X}^{[s]} = \mathbf{X}_{(\Pi^{[s]})}$ to be our current CPT copy of the original data vector $\mathbf{X}$, we track the log-likelihood, $\Sigma_{i=1}^{n} q(X_i^{[s]}|Z_i)$, and the correlation with the original data vector, corr$(\mathbf{X}, \mathbf{X}^{[s]})$. (Since X is strongly dependent on Z, it is to be expected that two draws of the data, i.e. $\mathbf{X}$ and $\mathbf{X}^{[s]}$, will necessarily have a high correlation.) The trace plots of these two quantities, plotted over $s = 0, 1, 2, \ldots, 250$ in Fig. 4, demonstrate that, in this simulation, the Markov chain appears to mix quickly, within about 50 or 100 iterations. Of course, this will be affected by factors such as the strength of the dependence between X and Z, and the sample size n.

6.3. *Capital Bikeshare data set*

We next implement the CPT and CRT on the Capital Bikeshare data set. (The data were obtained from `https://www.capitalbikeshare.com/system-data`.) Capital Bikeshare is a bike sharing programme in Washington DC, where users may check out a bike from one of their locations and return it at any other location. The data set contains each ride ever taken, recording the start time and location, end time and location, bike identification 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:

(a) test data set, all rides taken on weekdays (Monday–Friday) in October 2011, and sample size $n = 7346$ rides, after an initial screening step (details are given below);

Table 3. *p*-values obtained from the CPT and CRT for the Capital Bikeshare data†

Variable Y	*CPT p-value*	*CRT p-value*
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)

†The mean *p*-value and standard error (in parentheses) are calculated from 10 trials of each experiment (the randomness comes from the construction of the copies $\mathbf{X}^{(m)}$ for each test).

(b) training data set (for fitting the conditional distribution $Q(\cdot|Z)$), all rides taken on weekdays in September 2011 and November 2011, and sample size $n_{\text{train}} = 149912$ 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.

To implement the CPT and CRT, we shall use a conditional normal distribution, i.e. $(X|Z = z) \sim \mathcal{N}\{\mu(z), \sigma^2(z)\}$ as an estimate $Q(\cdot|z)$ of $Q_*(\cdot|z)$. 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 the CRT, we sample $M = 1000$ copies of $\mathbf{X}$ 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.

6.3.1. Results

We test the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ for several choices of the response Y.

(a) 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.
(b) 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.
(c) Day of the week (Monday–Friday), treated as categorical: bike riders' behaviour may differ on different days of the week, for instance, if rides on Friday are more likely to be leisure rides than on the other days of the week.

For user type and date, the statistic $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ that we use is the correlation between the vector $\mathbf{Y}$ 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 $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is given by

$$\max_{y \in \{\text{Mon},\ldots,\text{Fri}\}} \left| \text{correlation between } (R_1, \ldots, R_n) \text{ and } (\mathbb{1}(Y_1 = y), \ldots, \mathbb{1}(Y_n = y)) \right|.$$

Table 3 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; in contrast, 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 CPT that modifies the standard permutation test of independence between X and Y 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. In contrast, our theoretical understanding of type I error control covers the worst-case scenario over all possible statistics, and it may be that the simple statistics that are 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 to examine whether proposing swaps non-uniformly may improve the speed at which we can obtain copies $\mathbf{X}^{(m)}$ that are not too correlated with each other.

Furthermore, although in many applications we have access to plenty of unlabelled data, there will certainly be some domains where this is not so and it may not be possible to estimate the conditional distribution of $X|Z$ independently of the data. If only a small labelled data set (X, Y, Z) is available, with no additional unlabelled data (X, Z) with which to estimate this distribution, we would not want to split the data set to use one half for fitting $Q(X|Z)$ and the remaining half to run the CPT, since this would incur both a substantial loss in the type I error control (under the theoretical results of Section 5.1) and loss of 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 are used for estimating $Q(X|Z)$ and then reused for testing $H_0 : X \perp\!\!\!\perp Y|Z$. It is possible that tools from the selective inference literature may enable us to develop theory towards addressing this question.

Finally, both the CPT and the CRT are based on a setting where it is assumed that modelling $X|Z$ is easy whereas modelling $Y|X, Z$ is difficult—i.e. our estimate $Q(\cdot|Z)$ of the conditional distribution $X|Z$ is assumed to be highly accurate, but testing $H_0 : X \perp\!\!\!\perp Y|Z$ is a substantial challenge. In contrast, many of the asymptotic tests that were described in Section 2.3 treat the X- and Y-variables symmetrically when testing $X \perp\!\!\!\perp Y|Z$. 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?

Acknowledgements

RFB was partially supported by the National Science Foundation via grant DMS-1654076 and by an Alfred P. Sloan fellowship. TBB and RJS were supported by an Engineering and Physical Sciences Research Council programme grant. RJS was also supported by an Engineering and Physical Sciences Research Council Fellowship and a grant from the Leverhulme Trust. The authors thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the programme 'Statistical scalability' which was supported by Engineering and Physical Sciences Research Council grant LNAG/036, RG91310. The authors thank Samir Khan for help in

implementing code for our algorithms. We thank the reviewers for helpful and constructive feedback on an earlier draft.

Appendix A: Proofs

A.1. Proving validity of the sampling mechanisms

A.1.1. Proof of theorem 2

The proof of theorem 2 consists of simply checking the detailed balance equations for the Markov chain that is defined by the algorithm.

Let $\mathcal{P}$ be the set of all partitions of $\{1,\dots,n\}$ into $\lfloor n/2 \rfloor$ disjoint pairs. For any $p \in \mathcal{P}$ and any permutations π and π', we write $\pi \sim_p \pi'$ if π can be transformed to π' by swapping any subset of the pairs in the partition p. For example, if (i, j) and (k, l) are two of the disjoint pairs in the partition p, and π and π' are related via $\pi' = \pi \circ \sigma_{ij} \circ \sigma_{kl}$, then '$\pi \sim_p \pi'$ (recall that σ_{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 that is defined by algorithm 1. For ease of notation, for the remainder of this proof, we shall condition on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$ implicitly. In particular, all probabilities $\mathbb{P}(\cdot)$ or $\mathbb{P}(\cdot|\cdot)$ should be interpreted as $\mathbb{P}(\cdot|\mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z})$ or $\mathbb{P}(\cdot|\cdot, \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z})$.

For any permutations π and π', we have

$$\mathbb{P}(\Pi^{[t]} = \pi' | \Pi^{[t-1]} = \pi) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{P}(\Pi^{[t]} = \pi' | \Pi^{[t-1]} = \pi, t\text{th partition} = p),$$

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 that is defined for each $B_{t,k}$ in the odds ratio in Table 1, we see that, for any $\pi', \pi'' \sim_p \pi$,

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

where in the last step we refer to the distribution (5) of the permutation Π conditional on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$. Therefore,

$$\mathbb{P}(\Pi^{[t]} = \pi' | \Pi^{[t-1]} = \pi) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{\mathbb{1}(\pi' \sim_p \pi)\, \mathbb{P}(\Pi = \pi')}{\sum_{\pi''} \mathbb{1}(\pi'' \sim_p \pi)\, \mathbb{P}(\Pi = \pi'')}.$$

Thus, for any π and π', since '$\sim_p$' forms an equivalence relation over permutations, we have

$$\begin{aligned}
\mathbb{P}(\Pi = \pi)\, \mathbb{P}(\Pi^{[t]} = \pi' | \Pi^{[t-1]} = \pi) &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{P}(\Pi = \pi) \frac{\mathbb{P}(\pi' \sim_p \pi)\, \mathbb{P}(\Pi = \pi')}{\sum_{\pi''} \mathbb{P}(\pi'' \sim_p \pi)\, \mathbb{P}(\Pi = \pi'')} \\
&= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{P}(\Pi = \pi') \frac{\mathbb{1}(\pi \sim_p \pi')\, \mathbb{P}(\Pi = \pi)}{\sum_{\pi''} \mathbb{1}(\pi'' \sim_p \pi')\, \mathbb{P}(\Pi = \pi'')} \\
&= \mathbb{P}(\Pi = \pi')\, \mathbb{P}(\Pi^{[t]} = \pi | \Pi^{[t-1]} = \pi').
\end{aligned}$$

This verifies the detailed balance equations, and so the Markov chain is reversible and has stationary distribution given by expression (5). Finally, it is trivial to see that this Markov chain is aperiodic and irreducible when $q(x|z)$ is positive for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, and so, in this case, the stationary distribution is unique.

A.1.2. Proof of theorem 3

The proof of theorem 3 follows directly from the fact that the Markov chain that is defined in algorithm 1 is reversible, as shown in the proof of theorem 2. This means that, under H_0, the permutations $\Pi, \Pi_\sharp, \Pi^{(1)}, \dots, \Pi^{(M)}$ can equivalently be drawn as follows: first draw $\Pi_\sharp$ from distribution (5) conditionally on $\mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$; then draw $\Pi, \Pi^{(1)}, \dots, \Pi^{(M)}$ via $M+1$ independent runs of algorithm 1 for S steps initialized at $\Pi^{[0]} = \Pi_\sharp$. Thus $\Pi, \Pi^{(1)}, \dots, \Pi^{(M)}$ are IID conditionally on $\Pi_\sharp, \mathbf{X}_{()}, \mathbf{Y}$ and $\mathbf{Z}$, and are therefore exchangeable.

A.2. Proving robust type I error control

A.2.1. Proof of theorem 4

First we prove the result for the CRT. Let $\check{\mathbf{X}}$ be an additional copy drawn also from $Q(\cdot|\mathbf{Z})$, independently of $\mathbf{Y}$ and of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$. Then, since conditionally on $\mathbf{Y}$ and $\mathbf{Z}$ the copies $\mathbf{X}, \check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are independent, we have

$$\begin{aligned} d_{\mathrm{TV}}\{((\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z}), ((\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z})\} &= d_{\mathrm{TV}}\{(\mathbf{X}|\mathbf{Y}, \mathbf{Z}), (\check{\mathbf{X}}|\mathbf{Y}, \mathbf{Z})\} \\ &= d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}. \end{aligned}$$

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

$$A_\alpha := \left\{ (\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) : \frac{1 + \sum_{m=1}^M \mathbb{1}\{T(\mathbf{x}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geqslant T(\mathbf{x}, \mathbf{Y}, \mathbf{Z})\}}{1+M} \leqslant \alpha \right\},$$

i.e. the set where we would obtain a p-value $p \leqslant \alpha$. Then

$$\begin{aligned} \mathbb{P}(p \leqslant \alpha|\mathbf{Y}, \mathbf{Z}) &= \mathbb{P}\{(\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} \\ &\leqslant \mathbb{P}\{(\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} + d_{\mathrm{TV}}\{((\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z}), ((\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z})\} \\ &= \mathbb{P}\{(\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} + d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}. \end{aligned}$$

Finally, since $\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are clearly IID after conditioning on $\mathbf{Y}$ and $\mathbf{Z}$, and are therefore exchangeable, by definition of A_α we must have

$$\mathbb{P}\{(\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} \leqslant \alpha,$$

proving the desired bound for the CRT.

Next we turn to the CPT, for which the analysis is more complicated since the $\mathbf{X}^{(m)}$s depend on the observed values in the vector $\mathbf{X}$. We shall use the fact that

$$\text{for any } (U, V) \text{ and } (U', V'), \text{ if } (V|U=u) \overset{\mathrm{d}}{=} (V'|U'=u) \text{ for any } u, \text{ then } d_{\mathrm{TV}}\{(U, V), (U', V')\} = d_{\mathrm{TV}}(U, U'). \tag{9}$$

Let $\check{\mathbf{X}}$ be drawn from $Q(\cdot|\mathbf{Z})$, independently of $\mathbf{Y}$, and let $\check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)}$ be draws from the CPT when we sample from the values of $\check{\mathbf{X}}$ instead of $\mathbf{X}$, i.e., independently for each $m = 1, \ldots, M$, we draw

$$\check{\mathbf{X}}^{(m)} = \check{\mathbf{X}}_{(\check{\Pi}^{(m)})} \qquad \mathbb{P}(\check{\Pi}^{(m)} = \pi|\check{\mathbf{X}}_{()}, \mathbf{Y}, \mathbf{Z}) \propto q^n(\check{\mathbf{X}}_{(\pi)}|\mathbf{Z}),$$

where $\check{\mathbf{X}}_{()}$ and $\check{\mathbf{X}}_{(\pi)}$ are defined analogously to $\mathbf{X}_{()}$ and $\mathbf{X}_{(\pi)}$ from Section 3. Next, by comparing with the CPT sampling mechanism (6), we observe that the $\check{\mathbf{X}}^{(m)}$s, conditionally on $\check{\mathbf{X}}$, are generated with the same mechanism as the $\mathbf{X}^{(m)}$s conditionally on $\mathbf{X}$. In other words, for any $\mathbf{x} \in \mathcal{X}^n$, we have

$$((\check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)})|\check{\mathbf{X}} = \mathbf{x}, \mathbf{Y}, \mathbf{Z}) \overset{\mathrm{d}}{=} ((\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{X} = \mathbf{x}, \mathbf{Y}, \mathbf{Z}).$$

We can verify that the same equality in distribution holds if we instead use the exchangeable sampler (algorithm 2) with some choice $S \geqslant 1$ of the number of steps.

In either case, then, applying results (9) we have

$$\begin{aligned} d_{\mathrm{TV}}\{((\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z}), ((\check{\mathbf{X}}, \check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)})|\mathbf{Y}, \mathbf{Z})\} &= d_{\mathrm{TV}}\{(\mathbf{X}|\mathbf{Y}, \mathbf{Z}), (\check{\mathbf{X}}|\mathbf{Y}, \mathbf{Z})\} \\ &= d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}. \end{aligned}$$

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

$$\mathbb{P}(p \leqslant \alpha|\mathbf{Y}, \mathbf{Z}) \leqslant \mathbb{P}\{(\check{\mathbf{X}}, \check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} + d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\},$$

and, since $\check{\mathbf{X}}, \check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)}$ are exchangeable after conditioning on $\mathbf{Y}$ and $\mathbf{Z}$, we see that $\mathbb{P}\{(\check{\mathbf{X}}, \check{\mathbf{X}}^{(1)}, \ldots, \check{\mathbf{X}}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\} \leqslant \alpha$, proving the desired bound for the CPT (with permutations drawn either IID as in expression (6), or from the exchangeable sampler given in algorithm 2).

A.2.2. Proof of theorem 5

For convenience we shall write

$$d_{\mathrm{TV}} = d_{\mathrm{TV}}\{Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\}$$

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

$$\mathbb{P}_{Q^n_*(\cdot|\mathbf{Z})}\{\mathbf{X} \in A(\mathbf{Z})|\mathbf{Z}\} = \mathbb{P}_{Q^n(\cdot|\mathbf{Z})}\{\mathbf{X} \in A(\mathbf{Z})|\mathbf{Z}\} + d_{\mathrm{TV}}.$$

Fix any $M \geqslant 2$, and define

$$\alpha_0(\mathbf{Z}) := \mathbb{P}_{Q^n(\cdot|\mathbf{Z})}\{\mathbf{X} \in A(\mathbf{Z})|\mathbf{Z}\},$$
$$\alpha(\mathbf{Z}) := \alpha_0(\mathbf{Z}) + 0.5\sqrt{\left\{\frac{\log(M)}{M}\right\}}.$$

Now, by definition of the setting and the CRT, we know that, conditionally on $\mathbf{Z}$, we have $\mathbf{X} \sim Q^n_*(\cdot|\mathbf{Z})$ and, independently, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)} \sim Q^n(\cdot|\mathbf{Z})$. Therefore,

$$(\mathbb{1}\{\mathbf{X} \in A(\mathbf{Z}\}|\mathbf{Y}, \mathbf{Z}) \sim \text{Bernoulli}\{\alpha_0(\mathbf{Z}) + d_{\mathrm{TV}}\},$$

and, independently,

$$\left(\sum_{m=1}^{M} \mathbb{1}\{\mathbf{X}^{(m)} \in A(\mathbf{Z})\}|\mathbf{Y}, \mathbf{Z}\right) \sim \text{binomial}\{M, \alpha_0(\mathbf{Z})\}.$$

We shall work with the statistic $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \mathbb{1}\{\mathbf{X} \in A(\mathbf{Z})\}$. We have

$$\begin{aligned}
\mathbb{P}\{p \leqslant \alpha(\mathbf{Z})|\mathbf{Y}, \mathbf{Z}\} &= \mathbb{P}\left[\frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{X}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geqslant T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\}}{1+M} \leqslant \alpha(\mathbf{Z}) \middle| \mathbf{Y}, \mathbf{Z}\right] \\
&\geqslant \mathbb{P}\left[\mathbf{X} \in A(\mathbf{Z}) \text{ and } \sum_{m=1}^{M} \mathbb{1}\{\mathbf{X}^{(m)} \in A(\mathbf{Z})\} \leqslant \alpha(\mathbf{Z})(M+1) - 1 \middle| \mathbf{Y}, \mathbf{Z}\right] \\
&= \{\alpha_0(\mathbf{Z}) + d_{\mathrm{TV}}\}\,\mathbb{P}[\text{binomial}\{M, \alpha_0(\mathbf{Z})\} \leqslant \alpha(\mathbf{Z})(M+1) - 1|\mathbf{Z}] \\
&\geqslant \alpha(\mathbf{Z}) + d_{\mathrm{TV}} - 0.5\sqrt{\left\{\frac{\log(M)}{M}\right\}} - \mathbb{P}[\text{binomial}\{M, \alpha_0(\mathbf{Z})\} > \alpha(\mathbf{Z})(M+1) - 1|\mathbf{Z}],
\end{aligned} \tag{10}$$

where the last step holds by definition of $\alpha(\mathbf{Z})$ and $\alpha_0(\mathbf{Z})$, and the fact that $\alpha_0(\mathbf{Z}) + d_{\mathrm{TV}} \leqslant 1$. 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

$$\begin{aligned}
&\mathbb{P}\left(\text{binomial}(M, t) > \left[t + 0.5\sqrt{\left\{\frac{\log(M)}{M}\right\}}\right](M+1) - 1\right) \\
&\qquad = \mathbb{P}[\text{binomial}(M, t) - Mt > t + 0.5\sqrt{\left\{\frac{\log(M)}{M}\right\}}(M+1) - 1] \\
&\qquad \leqslant \exp\left(-Mt(1-t)\,h\left[\frac{t + 0.5\sqrt{\{\log(M)/M\}}(M+1) - 1}{Mt(1-t)}\right]\right) \\
&\qquad \leqslant \exp\left(-\frac{M}{4}h\left[\frac{0.5\sqrt{\{\log(M)/M\}}(M+1) - 1}{M/4}\right]\right),
\end{aligned} \tag{11}$$

where the last step holds since h is an increasing function, whereas $c \mapsto c\,h(a/c)$ is decreasing in $c > 0$, for any $a > 0$, and $t(1-t) \leqslant \frac{1}{4}$.

Finally, as $\epsilon \to 0$, we have $h(\epsilon) = \epsilon^2/2 + O(\epsilon^3)$, so as $M \to \infty$ we have

$$\begin{aligned}
\exp\left(-\frac{M}{4}h\left[\frac{0.5\sqrt{\{\log(M)/M\}}(M+1) - 1}{M/4}\right]\right) &= \exp\left\{-\frac{1}{2}\log(M) + o(1)\right\} \\
&= \frac{1}{\sqrt{M}} = 0.5o(1)\sqrt{\left\{\frac{\log(M)}{M}\right\}}.
\end{aligned}$$

Returning to inequality (11), we see that

$$\mathbb{P}\{p \leqslant \alpha(\mathbf{Z})|\mathbf{Y}, \mathbf{Z}\} \geqslant \alpha(\mathbf{Z}) + d_{\mathrm{TV}} - 0.5\sqrt{\left\{\frac{\log(M)}{M}\right\}}\{1 + o(1)\}.$$

More concretely, for any $M \geqslant 2$ we can verify numerically that the quantity in inequality (11) is bounded

by $2\sqrt{\{\log(M)/M\}}$, which shows that the term $0.5\{1+o(1)\}$ above can be replaced with 2.5 for any $M \geqslant 2$.

Appendix B: Details for the bike share data experiment

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

To estimate a conditional distribution $Q(\cdot|Z)$, we assume that $X|Z$ 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 $z = (z_{\text{route}}, z_{\text{time}})$,

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

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

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

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

for a bandwidth h of 20 min. Time of day is on a continuous 24-h clock, i.e. if $z_{\text{time}} = 11.00$ p.m. and $(Z_i^{\text{train}})_{\text{time}} = 1.00$ a.m. then the difference between them is 2 h, not 22 h.

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

$$(X|Z = z) \sim \mathcal{N}\{\hat{\mu}(z), \hat{\sigma}^2(z)\}.$$

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 because of 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_{route}, and at a time of day that was similar to z_{time}. For the test data, we then keep only those data points (X_i, Y_i, Z_i) for which $N(Z_i) \geqslant 20$. Since this screening step uses the value of Z_i but not the value of X_i, the X_is are still unobserved even after screening, and their distribution conditionally on Z_i is unchanged; therefore the CPT and CRT tests are valid even on these screened data.

References

Athey, S., Eckles, D. and Imbens, G. W. (2018) Exact *p*-values for network interference. *J. Am. Statist. Ass.*, **113**, 230–240.

Barber R. F. and Candès, E. (2019) On the construction of knockoffs in case–control studies. *Stat.*, **8**, no. 1, article e225.

Barber, R. F., Candès, E. J. and Samworth, R. J. (2019) Robust inference with knockoffs. *Ann. Statist.*, to be published.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, **81**, 608–650.

Bergsma, W. P. (2004) Testing conditional independence for continuous random variables. *Report 2004-048*. Eurandom, Eindhoven. (Available from `eurandom.tue.nl/reports/2004/048-report.pdf`.)

Berrett, T. B. and Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika*, to be published.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B, **80**, 551–577.

Cover, T. M. and Thomas, J. A. (2012) *Elements of Information Theory*. Chichester: Wiley.

Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B, **41**, 1–31.

Doran, G., Muandet, K., Zhang, K. and Schölkopf, B. (2014) A permutation-based kernel conditional independence test. *Uncertnty Artif. Intell.*, **30**, 132–141.
Ernst, M. D. (2004) Permutation methods: a basis for exact inference. *Statist. Sci.*, **19**, 676–685.
Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008) Kernel measures of conditional dependence. *Adv. Neurl Inform. Process. Syst.*, **20**, 489–496.
Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005) Measuring statistical dependence with Hilbert–Schmidt norms. In *Proc. 16th Int. Conf. Algorithmic Learning Theory* (eds S. Jain, H. U. Simon and E. Tomita), pp. 63–77. Berlin: Springer.
Hennessy, J., Dasgupta, T., Luke, M., Pattanayak, C. and Sarkar, P. (2016) A conditional randomization test to account for covariate imbalance in randomized experiments. *J. Causl Inf.*, **4**, 61–80.
Josse, J. and Holmes, S. (2013) Measures of dependence between random vectors and tests of independence: literature review. *Preprint arXiv:1307.7383*. École Polytechnique, Paris.
Kojadinovic, I. and Holmes, M. (2009) Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *J. Multiv. Anal.*, **100**, 1137–1154.
Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018) Kernel-based tests for joint independence. *J. R. Statist. Soc.* B, **80**, 5–31.
Roach, J. and Valdar, W. (2018) Permutation tests of non-exchangeable null models. *Preprint arXiv:1808.10483*. University of North Carolina at Chapel Hill, Chapel Hill.
Rosenbaum, P. R. (1984) Conditional permutation tests and the propensity score in observational studies. *J. Am. Statist. Ass.*, **79**, 565–574.
Runge, J. (2018) Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proc. 21st Int. Conf. Artificial Intelligence and Statistics* (eds A. Storkey and F. Perez-Cruz), pp. 938–947.
Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G. and Shakkottai, S. (2017) Model-powered conditional independence test. *Adv. Neurl Inform. Process Syst.*, **31**, 2955–2965.
Shah, R. D. and Peters, J. (2019) The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, to be published.
Song, K. (2009) Testing conditional independence via Rosenblatt transforms. *Ann. Statist.*, **37**, 4011–4045.
Stigler, S. M. (1989) Francis Galton's account of the invention of correlation. *Statist. Sci.*, **4**, 73–79.
Strobl, E. V., Zhang, K. and Visweswaran, S. (2019) Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causl Inf.*, **7**, no. 1.
Su, L. and White, H. (2007) A consistent characteristic function-based test for conditional independence. *J. Econmetr.*, **141**, 807–834.
Su, L. and White, H. (2008) A nonparametric Hellinger metric test for conditional independence. *Econmetr. Theory*, **24**, 829–864.
Su, L. and White, H. (2014) Testing conditional independence via empirical likelihood. *J. Econmetr.*, **182**, 27–44.
Székely, G. J. and Rizzo, M. L. (2014) Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, **42**, 2382–2412.
Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.
Veraverbeke, N., Omelka, M. and Gijbels, I. (2011) Estimation of a conditional copula and association measures. *Scand. J. Statist.*, **38**, 766–780.
Weihs, L., Drton, M. and Meinshausen, N. (2018) Symmetric rank covariances: a generalised framework for nonparametric measures of dependence. *Biometrika*, **105**, 547–562.
Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2011) Kernel-based conditional independence test and application in causal discovery. *Uncertnty Artif. Intell.*, **27**, 804–813.